COVID Impact Tracking and Clustering

Chintan Desai

March 31st , 2021

1. Business Problem & Solution:

1.1 Problem Statement:

With the increase of COVID disease, it is critical to impose strict isolation rules in severely impacted regions. However, as the disease is not expected to go away soon, it is also very important to avoid adverse impact on economy due to lockdown for a longer period. Regions with severe impact also cannot follow complete shutdown and hence it is crucial to understand which business/venues/places are more prone to spread of the disease and only those needs to follow strict rules, while the rest can observe little lighter rules to keep economy running and help reducing impact on day-to-day life of local residents.

At present we have several dashboards which shows location wise impact of the disease, but none of them provides view on what are the most common business/venues in those locations, which play vital role in determining strategy to impose isolation and/or lockdown rules. Getting this full view on most common business/services around various regions with severity of the disease in those regions is challenging in absence of such dashboard.

1.2 Goal Statement:

In this project, I primarily aim to -

- Identify one of the most COVID impacted place of the world
- Explore various regions of that place
- Cluster them together based on type of most common venues/businesses in surrounding area
- Create visualizations to display region wise severity of the disease along with most common venues/businesses in those regions

Achieving above goal will help our target audience to determine strategy to impose isolation and/or lockdown rules in each region and thus control the spread of the disease and minimize impact on economy.

1.3 Target Audience:

Who can benefit from this project outcome?

- Local authorities and government of the severely impacted county/city to determine the best strategy to impose isolation/lockdown rules
- Local Police department to get strict isolation rules followed by public
- Local residents to understand severity of the disease by regions and venues/types of business

2. Data Acquisition & Cleaning:

2.1 Data Sources and API/Libraries:

In order to perform my analysis, I required various data including COVID cases statistics, location coordinates of various parts of the world, surrounding venues of the COVID impacted regions..etc. I collected this information using following data sources.

Data was collected from various sources as follows -

- **WHO**: Country wise Total # of COVID cases across the world
- **Worldometer-USA:** State wise Total # of COVID cases across USA
- **Worldometer-California:** County wise Toal # of COVID cases across state of California
- **LA county public Dashboard:** Region wise Total # of COVID cases across Los Angeles county
- **The Geocoder Python library:** Location Coordinates of various regions of LA county
- **Folium Python library:** Visualize geographic details of region or place
- **Foursquare API:** Explore various regions and find out most common venues around those regions to cluster them together
- **Choropleth Map:** Visualize COVID severity across world/country/state/county and display clusters of most common venues superimposed on covid severity map

2.2 Data Cleaning & Pre-processing:

Data that was downloaded or scrapped from websites was converted to pandas dataframe and columns were massaged to look more meaningful with shorter names and unnecessary columns were removed. From COVID count datasets, additional summary or title records were removed. After that any record with missing or null value in key fields like New Cases, Population..etc. was either removed or null value was replaced with 0. For certain records, Total number of cases information was not available and hence I excluded those records from the dataset. Also, all the number columns were converted to integer for my analysis.

In our dataset for Los Angeles county, urban and rural area were listed separately in LA county portal COVID dashboard. I combined those to a single record and aggregated population as well as COVID case counts. I had information on total population and number of cases for each county, so I calculated case percentage for each county and added back to my dataframe to use in further analysis.

3. Exploratory Data Analysis

3.1 Approach:

I utilized above data sources to perform my analysis and solved the business problem described above. Below is the high-level approach to summarize the flow of analysis I performed.

First step is to identify location in the world which is one of the most impacted places by COVID

- Collect required data from various sources mentioned above and preprocess the data aligning with the need
- Visualize COVID impact on various countries on world map on factors like total # of cases, spread per 100k population, total cases in last 7 days..etc.. using Choropleth
- Visualize state wise impact of the country that is severely impacted in most aspects
- Pick the most adversely impacted country in most aspects and visualize state wise impact of the disease on that country using choropleth
- Pick the most impacted state and visualize county wise impact of the disease on state using choropleth
- Pick the most impacted county and visualize various regions of the county to understand severity of the disease

Once most impacted county is identified, follow the next step to determine lockdown strategy based on approach described below.

- **Get** coordinates of county regions using Geocoder Python library
- **Explore** nearby venues of each of the region using Foursquare API
- **Analyze** the data using K-means clustering and superimpose clusters on the choropleth map showing COVID impact by various regions of the county
- **Categorize** each cluster based on top 3 most common venue and determine cluster wise COVID impact
- From the results, **conclude** severely, moderately and lightly impacted clusters and corresponding venue types
- **Recommend** lockdown strategy from conclusion of analysis

3.2 Data Visualization:

After data cleaning and pre-processing, I created visualizations using choropleth map to showcase COVID impact severity in various parts of the world. Some of the countries had a greater number of total cases due to their higher population, however when I compared countries by total #of cases per 100k population, those were not on the top of the list. However, USA was the only country which was leading the list in every category, and hence was most severely impacted country of the world by COVID

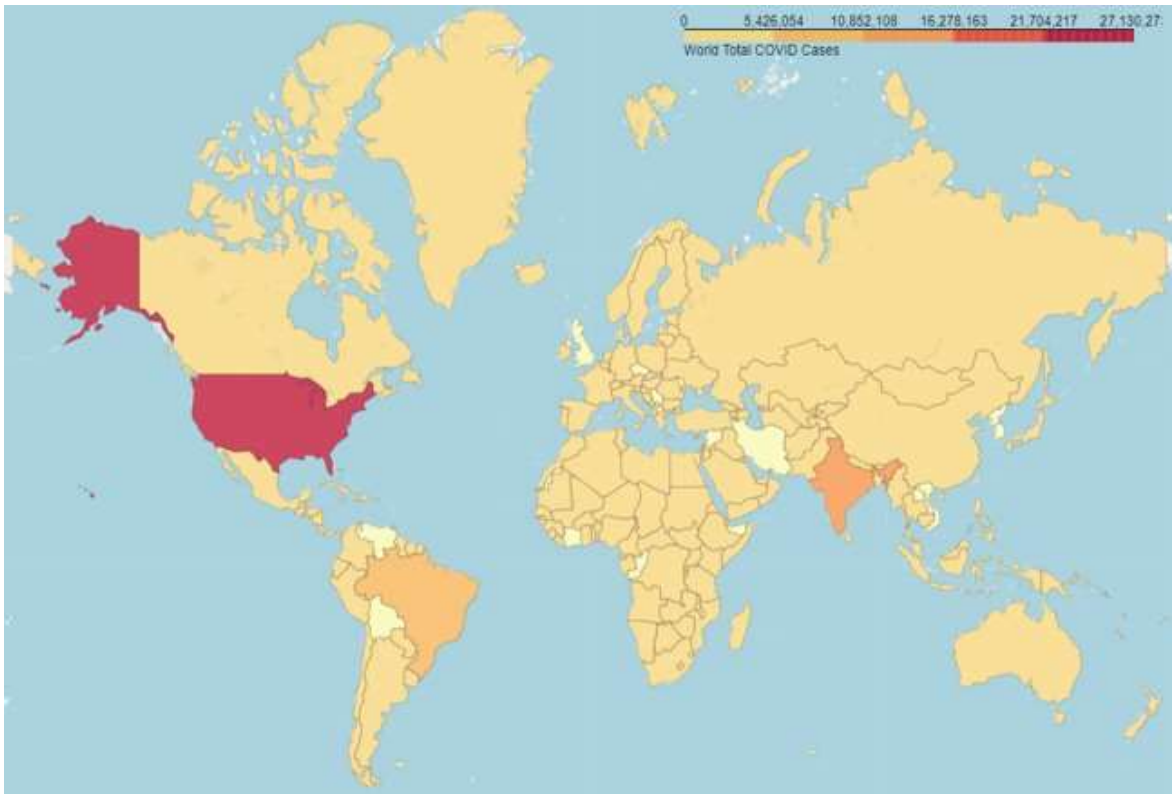

Figure 1. Impact of COVID on various countries in the world

As USA was the most impacted country, I further analyzed data across various USA states and found out California as one of the most impacted state to analyze further.

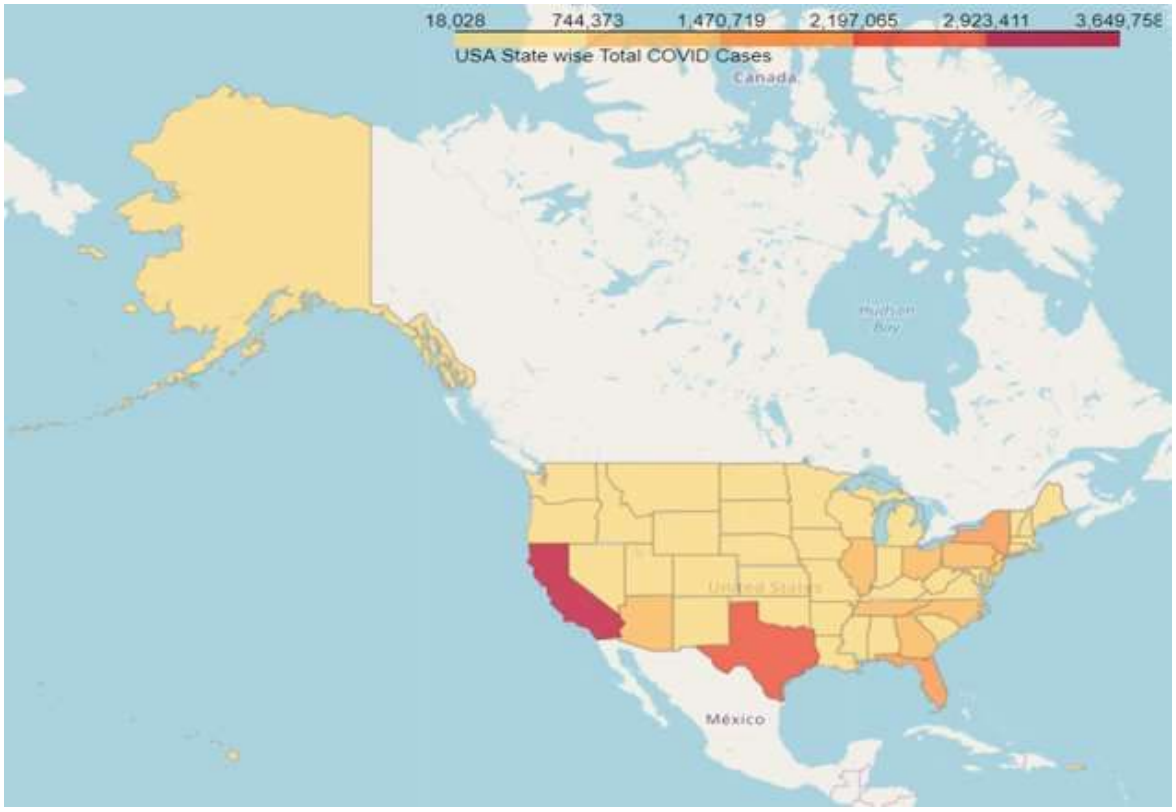

Figure 2. Impact of COVID on various states in USA

My next goal was to identify the most adversely impacted county of California state, which I can further explore for my analysis. With visualization shown below, I concluded Los Angeles was most adversely hit by COVID in California state and hence I decided to explore various region of LA to determine lockdown strategy in this county.

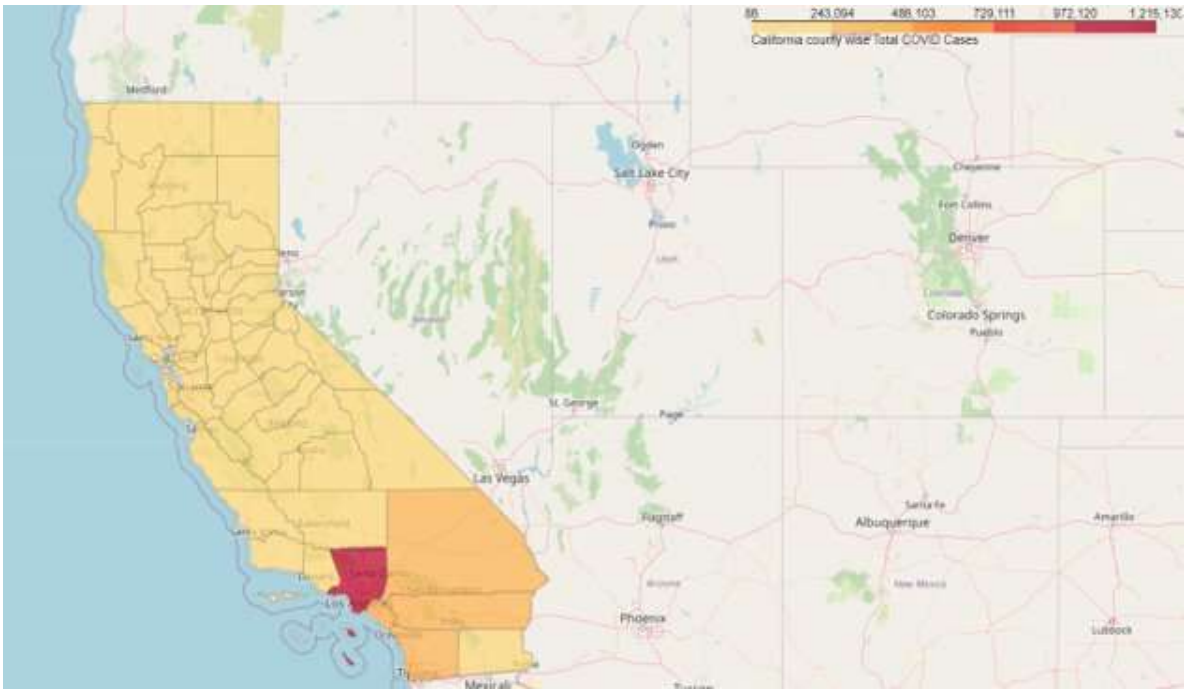

Figure 3. Impact of COVID on various county of California State

Region	index	Total_cases	population	Case_Percentage	Latitude	Longitude
East Los Angeles	60	24217	125269	19.33	34.03347	-118.159090
Pomona	313	23575	157869	14.93	34.05499	-117.750040
Palmdale	291	23353	159810	14.61	34.57936	-118.116590
Florence-Firestone	147	22051	112150	19.66	33.97475	-118.249932
Lancaster	118	20662	161570	12.79	34.69893	-118.144780
North Hollywood	140	19454	151421	12.85	34.16982	-118.378990
Santa Clarita	174	18543	220424	8.41	34.41389	-118.551180
Glendale	77	18091	206493	8.76	34.14633	-118.248640
South Gate	183	17892	98155	18.23	33.95722	-118.205630
Boyle Heights	24	16928	86884	19.48	34.04004	-118.210500

Figure 4. Top county list of California by total # of COVID cases

Below is the map created to visualize various regions of LA county.

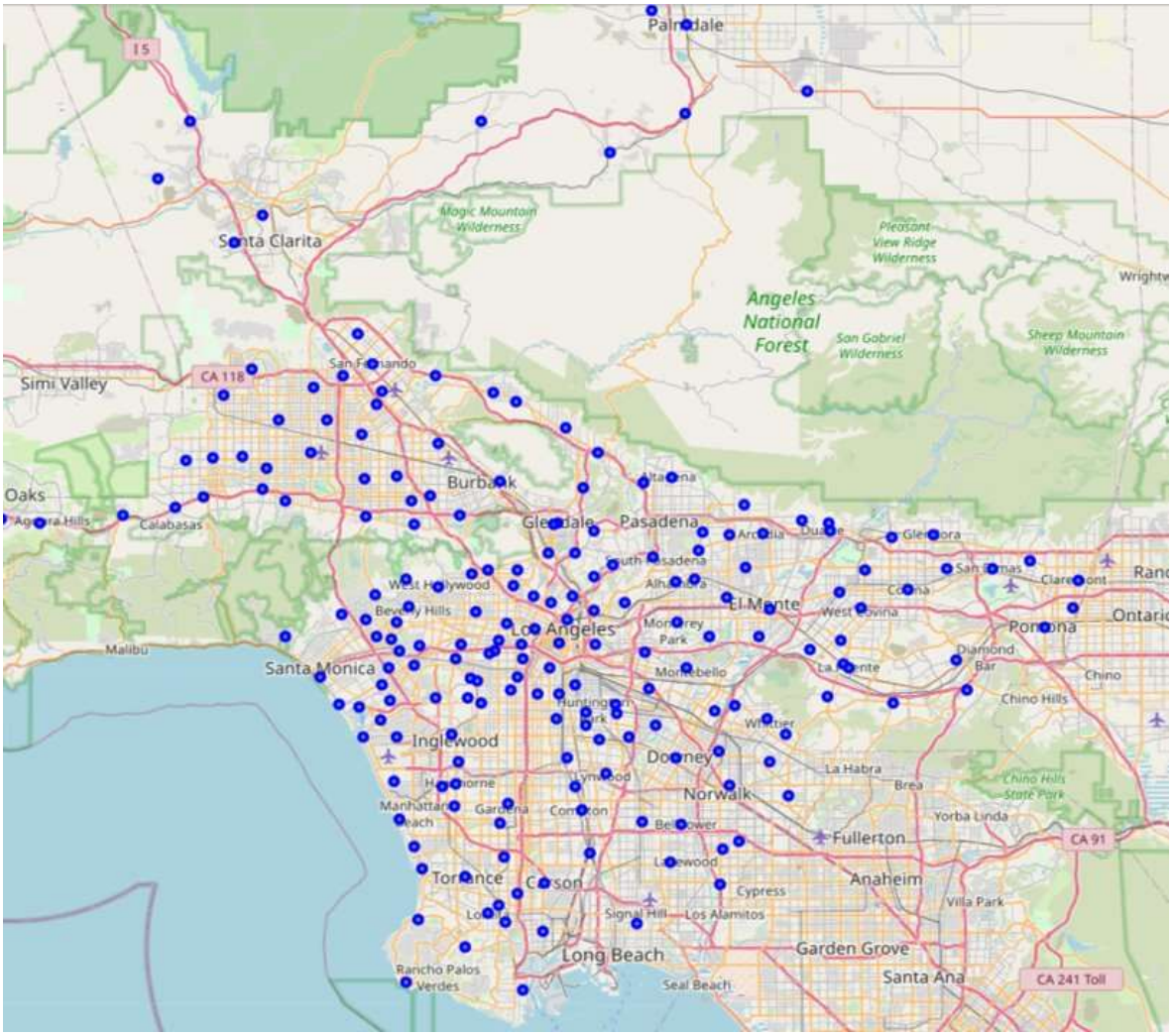

Figure 5. Visualize various regions on the map of LA County

3.3 Explore and Analyze each LA Neighborhood Region:

As I moved further, I retrieved location coordinate of each region of LA county using python geocoder library. I explored nearby 100 venues and venue categories within 1500 meter range for each county region using Foursquare API. I enriched my dataframe with location coordinates and nearby venues along with venue category as shown below -

Region	Region Latitude	Region Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Acton	34.46815	-118.19513	Acton Market & Country Store	34.468595	-118.197626	Grocery Store
Acton	34.46815	-118.19513	Fox Hay Feed and Grain	34.469565	-118.195481	Pet Store
Acton	34.46815	-118.19513	High Mesa	34.467814	-118.196090	Food
Acton	34.46815	-118.19513	Acton Market	34.467628	-118.195892	Grocery Store
Acton	34.46815	-118.19513	TSW Social Media Marketing	34.470898	-118.192307	Market

Figure 6. Dataframe enriched with location coordinates and venue details

Total number of regions I retrieved was 200 plus and hence number of nearby venues was also very large. Hence to reduce total number of distinct venue category for plotting, I put similar venue categories in a single venue category type as shown in example below -

Region	Region Latitude	Region Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Category Type
Acton	34.46815	-118.19513	Acton Market & Country Store	34.468595	-118.197626	Grocery Store	Grocery
Acton	34.46815	-118.19513	Fox Hay Feed and Grain	34.469565	-118.195481	Pet Store	Public Service Place
Acton	34.46815	-118.19513	High Mesa	34.467814	-118.196090	Food	Food
Acton	34.46815	-118.19513	Acton Market	34.467628	-118.195892	Grocery Store	Grocery
Acton	34.46815	-118.19513	TSW Social Media Marketing	34.470898	-118.192307	Market	Grocery

Figure 7. Venue Category Type added to consolidate multiple categories into one

After that, I found frequency of given venue type in each region to understand what are the most common venues in each region.

Region	Accomodation	American Restaurant	Art/Craft/Flower Shop	Art/Museum	Asian Restaurant	Australian Restaurant	Bakery/Breakfast Spot	Bank/ATM	Bar/Pub/Liquor Store	...	Pizza/Salad Place	Public Entertainment Place	Public Service Place
Acton	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.000000	...	0.000000	0.0	0.166667
Adams-Normandie	0.000000	0.153846	0.076923	0.000000	0.000000	0.0	0.000000	0.0	0.000000	...	0.076923	0.0	0.000000
Agoura Hills	0.095238	0.047619	0.047619	0.000000	0.047619	0.0	0.000000	0.0	0.095238	...	0.047619	0.0	0.047619
Agua Dulce	0.000000	0.111111	0.000000	0.000000	0.000000	0.0	0.111111	0.0	0.000000	...	0.111111	0.0	0.000000
Alhambra	0.000000	0.022727	0.000000	0.022727	0.159091	0.0	0.068182	0.0	0.045455	...	0.022727	0.0	0.045455

```
----Acton----
                        venue  freq
0                     Grocery  0.50
1                        Food  0.17
2          Office/Business Place  0.17
3          Public Service Place  0.17
4                 Accomodation  0.00
5                    Pharmacy  0.00
6            Mexican Restaurant  0.00
7  Miscellaneous Shopping Place  0.00
8                Movie/Theater  0.00
9            Park/Garden/Beach  0.00

----Adams-Normandie----
                        venue  freq
0          American Restaurant  0.15
1                         Food  0.15
2                Gym/Yoga/Spa  0.08
3          Art/Craft/Flower Shop  0.08
4                       Sports  0.08
5                      Grocery  0.08
6               Sandwich Place  0.08
7           Park/Garden/Beach  0.08
8                  Coffee Shop  0.08
9           Pizza/Salad Place  0.08

----Agoura Hills----
                        venue  freq
0      Furniture / Home Service  0.14
1                 Accomodation  0.10
2          Bar/Pub/Liquor Store  0.10
3          Fast Food Restaurant  0.05
4          Public Service Place  0.05
5             Pizza/Salad Place  0.05
6  Miscellaneous Shopping Place  0.05
7            Italian Restaurant  0.05
8          American Restaurant  0.05
9                  Gym/Yoga/Spa  0.05
```

Figure 8. Frequency of Venue Category Type for each region of LA county

Based on this, I enriched my dataframe with top 10 most common venues for each region.

Region	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
Acton	Grocery	Public Service Place	Office/Business Place	Food	Art/Museum	Asian Restaurant	Australian Restaurant	Art/Craft/Flower Shop	Bakery/Breakfast Spot	Bank/ATM
Adams-Normandie	Food	American Restaurant	Pizza/Salad Place	Sports	Gym/Yoga/Spa	Coffee Shop	Grocery	Sandwich Place	Art/Craft/Flower Shop	Park/Garden/Beach
Agoura Hills	Furniture / Home Service	Accomodation	Bar/Pub/Liquor Store	Art/Craft/Flower Shop	Fast Food Restaurant	American Restaurant	Gym/Yoga/Spa	Italian Restaurant	Miscellaneous Shopping Place	Cafe
Agua Dulce	Grocery	American Restaurant	Electronics Store	Mexican Restaurant	Bakery/Breakfast Spot	Pizza/Salad Place	Cafe	Clothing Store	Fast Food Restaurant	European Restaurant
Alhambra	Asian Restaurant	Grocery	Dessert/Ice Cream Shop	Bakery/Breakfast Spot	Sandwich Place	Bar/Pub/Liquor Store	Public Service Place	Seafood Restaurant	Cafe	American Restaurant
Altadena	Grocery	Food	Miscellaneous Shopping Place	Pizza/Salad Place	Electronics Store	Furniture / Home Service	Coffee Shop	Mexican Restaurant	Bar/Pub/Liquor Store	Bank/ATM
Arcadia	Park/Garden/Beach	Miscellaneous Shopping Place	Bakery/Breakfast Spot	Vegetarian / Vegan Restaurant	Fast Food Restaurant	European Restaurant	Electronics Store	Dessert/Ice Cream Shop	Coffee Shop	Clothing Store
Arleta	Sports	Art/Craft/Flower Shop	Sandwich Place	Cafe	Food	Fast Food Restaurant	European Restaurant	Electronics Store	Dessert/Ice Cream Shop	Coffee Shop
Artesia	Asian Restaurant	Grocery	Dessert/Ice Cream Shop	Seafood Restaurant	Art/Craft/Flower Shop	Food	Miscellaneous Shopping Place	American Restaurant	Bakery/Breakfast Spot	Clothing Store
Atwater Village	American Restaurant	Asian Restaurant	Miscellaneous Shopping Place	Bar/Pub/Liquor Store	Coffee Shop	Pizza/Salad Place	Mediterranean Restaurant	Sandwich Place	Art/Craft/Flower Shop	Public Service Place

Figure 9. Top 10 most common venues nearby each region of LA county

3.4 Clustering LA county Regions:

I created clusters of LA county regions based on nearby venues for each region. I used K-means clustering and found out optimum value of k using Elbow method.

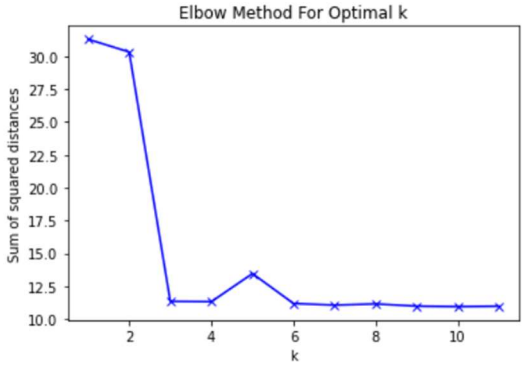

Figure 10. Optimum value of k

As seen above, value of k dropped drastically between 2 & 3 and then again increased. At k=6 it stabilized, so I kept number of clusters = 6 for my analysis.

I labelled each cluster according to frequency of most common venues in the cluster and created a result dataframe showing COVID severity in each cluster as shown below.

Region	index	Total_cases	population	Case_Percentage	Latitude	Longitude	Labels
Acton	0	425	7971	5.33	34.468150	-118.195130	2
Adams-Normandie	1	1129	8202	13.76	33.901212	-118.299321	3
Agoura Hills	2	930	20883	4.45	34.146110	-118.778120	3
Agua Dulce	3	246	4158	5.92	34.495700	-118.326210	2
Alhambra	4	6544	86724	7.55	34.094420	-118.127780	1

Figure 11. Cluster labels for each region

Labels	0	1	2	3	4	5
Total_cases	3966.461538	5440.966102	4736.363636	4390.061947	537.00	2413.25
population	31815.000000	49211.305085	30971.590909	41132.902655	8624.00	25956.75
Case_Percentage	11.961538	10.645932	12.121364	9.648230	32.69	6.29

Figure 12. Cluster wise COVID severity

3.5 Results:

I created a result dataframe showing cluster label for each region along with COVID case information as well as Top 10 most common value for each region of LA county.

Cluster Labels	Region	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
2	Acton	Grocery	Public Service Place	Office/Business Place	Food	Art/Museum	Asian Restaurant	Australian Restaurant	Art/Craft/Flower Shop	Bakery/Breakfast Spot	Bank/ATM
3	Adams-Normandie	Food	American Restaurant	Pizza/Salad Place	Sports	Gym/Yoga/Spa	Coffee Shop	Grocery	Sandwich Place	Art/Craft/Flower Shop	Park/Garden/Beach
3	Agoura Hills	Furniture / Home Service	Accomodation	Bar/Pub/Liquor Store	Art/Craft/Flower Shop	Fast Food Restaurant	American Restaurant	Gym/Yoga/Spa	Italian Restaurant	Miscellaneous Shopping Place	Cafe
2	Agua Dulce	Grocery	American Restaurant	Electronics Store	Mexican Restaurant	Bakery/Breakfast Spot	Pizza/Salad Place	Cafe	Clothing Store	Fast Food Restaurant	European Restaurant
1	Alhambra	Asian Restaurant	Grocery	Dessert/Ice Cream Shop	Bakery/Breakfast Spot	Sandwich Place	Bar/Pub/Liquor Store	Public Service Place	Seafood Restaurant	Cafe	American Restaurant

Figure 13. Dataframe with Cluster labels, venues and COVID cases

As we had multiple venues around each region, I used first three most common venues around every region and used those for clustering. I created a dataframe showing frequency of top 3 venues in each cluster as shown below -

Venue Category Type	Accomodation	American Restaurant	Art/Craft/Flower Shop	Art/Museum	Asian Restaurant	Bakery/Breakfast Spot	Bank/ATM	Bar/Pub/Liquor Store	Cafe	Clothing Store	...	Pharmacy	Pizza/Salad Place	Public Entertainment Place	Public Service Place
0	0	0	0	1	0	1	0	0	2	1	...	0	1	0	1
1	3	17	1	0	53	22	5	1	1	1	...	0	2	2	4
2	1	5	1	0	3	3	2	2	1	0	...	1	1	0	0
3	6	36	14	3	16	19	6	40	7	10	...	3	5	1	6
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	2
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0

Figure 14. Frequency of top3 most common venues in each cluster

In order to visualize frequency of top 3 most common venues category types in each cluster, I plotted the data in bar chart as shown below.

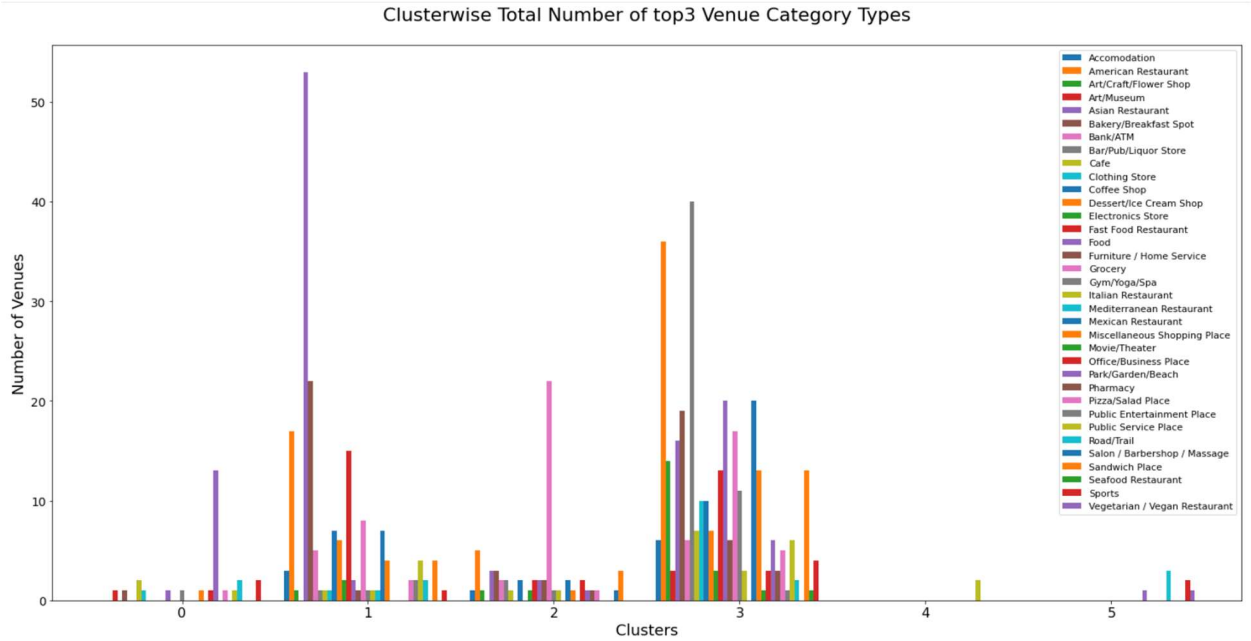

Figure 15. Visualization in Bar Chart for top3 venue category types frequency

Based on above chart, I derived clusters into following distinct Titles and displayed 3 distinct venue category types for each region-

Clusters	Cluster Title
0	Park/Beach Cluster
1	Asian Restaurant Cluster
2	Grocery Cluster
3	Bar & Food Cluster
4	Public Service Place Cluster
5	Road/Trail Cluster

Region	Venue Types
Acton	3 Grocery, 1 Food, 1 Office/Business Place
Adams-Normandie	2 American Restaurant, 2 Food, 1 Art/Craft/Flo...
Agoura Hills	3 Furniture / Home Service, 2 Accomodation, 2 ...
Agua Dulce	2 Grocery, 1 American Restaurant, 1 Bakery/Bre...
Alhambra	7 Asian Restaurant, 4 Grocery, 3 Bakery/Breakf...
Altadena	2 Food, 2 Grocery, 2 Miscellaneous Shopping Place
Arcadia	1 Bakery/Breakfast Spot, 1 Miscellaneous Shopp...
Arleta	1 Art/Craft/Flower Shop, 1 Sandwich Place, 1 S...
Artesia	22 Asian Restaurant, 4 Grocery, 3 Dessert/Ice ...
Atwater Village	4 American Restaurant, 3 Asian Restaurant, 3 B...

Figure 16. Cluster Title and top3 venues for each region

Finally, I superimposed clusters into LA county choropleth map showing COVID severity in various regions of the county.

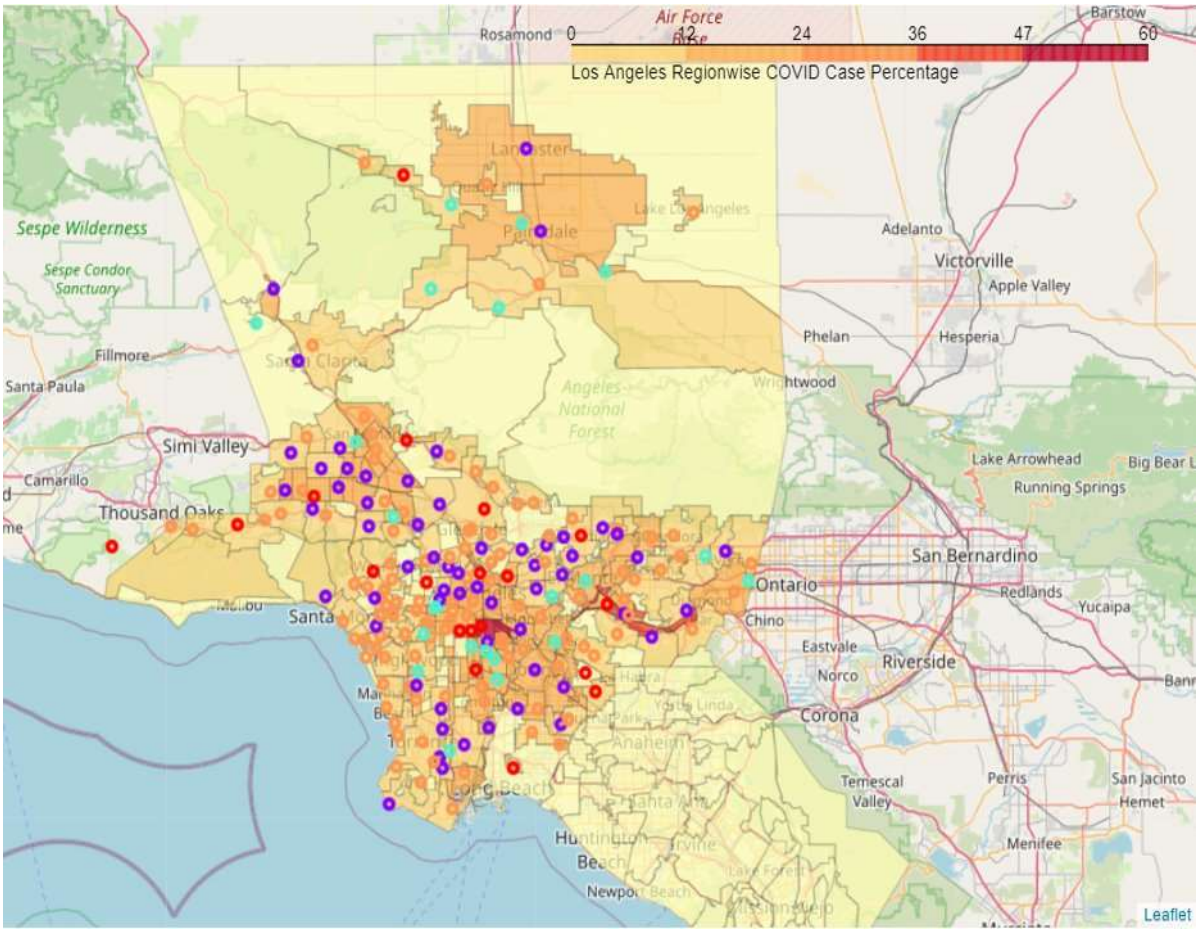

Figure 17. Clusters superimposed on COVID severity map of LA county

From above study, here is our final result table for each cluster showing COVID severity. This table is useful to understand impact of COVID in various clusters and determine lockdown strategy.

Clusters	Cluster Title	Total_cases	population	Case_Percentage
0	Park/Beach Cluster	3966.461538	31815.000000	11.961538
1	Asian Restaurant Cluster	5440.966102	49211.305085	10.645932
2	Grocery Cluster	4736.363636	30971.590909	12.121364
3	Bar & Food Cluster	4390.061947	41132.902655	9.648230
4	Public Service Place Cluster	537.000000	8624.000000	32.690000
5	Road/Trail Cluster	2413.250000	25956.750000	6.290000

Figure 18. Result table showing Clusters with title and COVID case statistics

4. Discussions and Recommendations

I have clustered LA county neighborhood regions based upon top 3 most common surrounding venues and superimposed those clusters on LA county map showing impacted population across various regions. This is very helpful to understand severely impacted regions and most common venues around those regions. Similar study can be extended to other locations of the world.

In above study, we could see that elbow point for k values is at 3, which again rises and then stabilizes at the value of 6. Hence, I have used k=6 (number of clusters) for this analysis. In study above, I have used canberra method to find the distance between two points. This data can be analyzed using other methods as well.

Additionally, I have used static data to pull COVID cases information in the world, LA county list as well as COVID cases information from LA county portal. This data can be fetched dynamically in future studies to retrieve up to date information at any point of time. Also, due to higher size of value categories, I defined venue category type and placed similar venues together in a single bucket in my analysis. In order to get more detailed view of analysis, this analysis can be extended at more granular level (without bucketing multiple venue into one) which will give more detailed results of COVID impact.

5. Conclusion

My aim for this study was to identify most common venues around one of the most severely impacted regions by COVID disease in the world. In my initial study, I identified USA as the most impacted country in all aspects. After that, I identified California as one of the most severely impacted and most populated state of USA. Hence, I focused my analysis on this state and further identified Los Angeles as most impacted county of California state.

After that, I explored Los Angeles county further and identified most common venues across various neighborhood regions of this county. Later I used K-means clustering to create clusters of top three most common venues category types around each region. Finally, I superimposed those clusters on LA county choropleth map, reflecting severity of COVID in terms of percentage of impacted population.

I concluded my analysis with the following results for each cluster -

Clusters	Cluster Title	Total_cases	population	Case_Percentage
0	Park/Beach Cluster	3966.461538	31815.000000	11.961538
1	Asian Restaurant Cluster	5440.966102	49211.305085	10.645932
2	Grocery Cluster	4736.363636	30971.590909	12.121364
3	Bar & Food Cluster	4390.061947	41132.902655	9.648230
4	Public Service Place Cluster	537.000000	8624.000000	32.690000
5	Road/Trail Cluster	2413.250000	25956.750000	6.290000

With above study, I am able to conclude that -

- USA is the most impacted country of the world by COVID disease in all aspects (total cases, Number of cases by 100k population, cases in last 7 days..etc)
- California, being the most populated state, is one of the most adversely impacted state of USA by COVID
- Los Angeles is the most impacted county of California, and hence should be focused for further analysis
- Most parts of Los Angeles have mixture of various social venues, mostly led by various continental restaurants (Asian, American & Mexican restaurants to be precise)
- **Cluster 4** is the least populated cluster and unlike other clusters, this doesn't have many social venues. Here most venues are primarily public service places (like Gas Station, Public transport station, warehouse, laundry). Although total number of cases are less here due to less population, in terms of impacted population percentage, it is the worst impacted cluster and hence needs to follow very strict isolation rules
- **Cluster 2** with most grocery places and **Cluster 1** with Asian and other Restaurants are in populated regions and also requires strict isolation rules to reduce the spread
- **Cluster 0** has more public gathering places like parks and beaches and **Cluster 3** has more bars/pubs and food places. These two clusters are also candidates for moderately strict regulation rules but can have necessary businesses open with precautions
- **Cluster 5** is the least impacted cluster and even though being crowded place, it does not have many social venues nearby except a few Trails and Sport shops and hence least impacted by COVID. This can follow comparatively lighter rules

This study will help our target audience to review the impact of COVID across various regions of LA and decide the best strategy to impose isolation rules around various regions and/or types of business.

6. References

Some of the key references used for this study are mentioned below.

[1]. WHO COVID Dashboard

[2]. Worldometer COVID Dashboard

[3]. LA County COVID Dashboard

[4]. LA County Regions Location Coordinates

[5]. Foursquare API