COVID 19 – Precautions and Venue Data Analysis of Seattle

Sathya Thiruvengadam

Jun 2020

Introduction

- The COVID-19 Pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019(COVID-19), caused by severe acute respiratory syndrome coronavirus 2(SARS-CoV-2). The virus that causes COVID-19 is spreading very easily and sustainably between people.
- The virus is spreading between person-to-person due to close contact, tiny droplets from sneezing, coughing can spread in air and people get infected by touching the contaminated areas and touching their face.
- As of today, the graph shows the consistent upward arch while in the other end thing are getting to normal called “new” normal with precautions. But still people find it difficult to get adjusted to new normal, where bringing more risk prone zones for easy contamination.

Problems and Interest

- By analyzing the venues of high risky area(where it is highly crowded) and making a legally bonded procedures with extra precautions to handle based on the nature of the venue is an optimal way to control the contamination while we step into the stage of introducing new normal to the people. The idea behind analyzing the venues of Seattle is due to the constant increase in COVID-19 cases and a big city with large crowd, in future this model can be used in any cities for this purpose as states started slowly announcing to reverse the lock down.
- Government official to consider extra measures to take precautions and monitor the process to minimize the contamination. General public knowing the high-risk prone areas to undertake extra personnel precautions or to the max avoiding the risk prone areas.

Data Description

- **Neighborhood Data :-**

  The data of Seattle Neighborhood is scraped from Wikipedia using BeautifulSoup- python library for web scraping.

- **Geocoding**

  Performing geocoding in python with the help of Geopy and Geopandas libraries to retrieve the **geographical coordinates** of an address from neighborhood data.

- **Venues for clustering**

  Using the geographical coordinates obtained, the venues are gathered using **FourSquare API** to perform clustering.

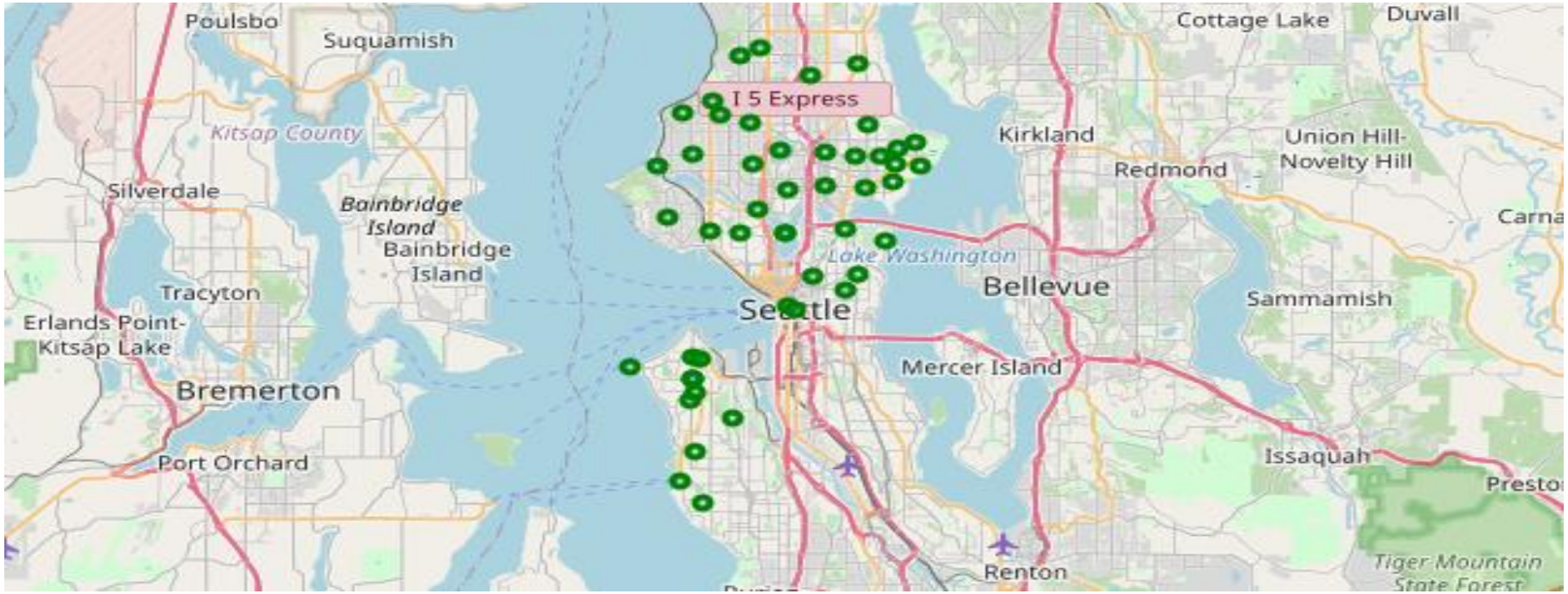
Methodology

Scraped data from Wikipedia, cleaned and used geocoder to obtain the geographical coordinates of the address listed, the data used below

	Neighborhood	level_1	Borough	Address	latitude	longitude
0	North Seattle	0	Seattle	North Seattle, Seattle	47.660773	-122.291497
1	Broadview	0	North Seattle	Broadview, North Seattle	47.722320	-122.360407
2	Bitter Lake	0	North Seattle	Bitter Lake, North Seattle	47.726236	-122.348764
3	North Beach	0	North Seattle	North Beach , North Seattle	47.696210	-122.392362
4	Blue Ridge	0	North Seattle	Blue Ridge, North Seattle	47.701487	-122.375407

Methodology

Used python **folium** library to visualize geographic details of Seattle and its boroughs and I created a map of Seattle with boroughs superimposed on top.



Methodology

Used the Foursquare API to explore the boroughs and segment them, set the boundaries with the radius of 1000 meters and allocated the limit as 50 venues for each borough from their geographical coordinates' location.

	Borough	Neighborhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Seattle	North Seattle	47.660773	-122.291497	Burke-Gilman Brewing Company	47.661308	-122.288067	Brewery
1	Seattle	North Seattle	47.660773	-122.291497	Jak's Grill	47.661072	-122.288073	Steakhouse
2	Seattle	North Seattle	47.660773	-122.291497	Center for Urban Horticulture	47.657978	-122.290237	College Science Building
3	Seattle	North Seattle	47.660773	-122.291497	University Village	47.662487	-122.298531	Shopping Plaza
4	Seattle	North Seattle	47.660773	-122.291497	The North Face	47.662400	-122.298158	Sporting Goods Shop

Methodology

From the detail count, it shows there are **269 unique categories** listed and now each neighborhood is analyzed individually to understand the most common place tend to be crowded within 1000 meters.

The process is taken further by using "**one-hot encoding**" function of the python panda's library'. One hot encoding converts the categorical variables into a form that could be help ML algorithms for better prediction accuracy.

	Zoo	Zoo Exhibit	Borough	ATM	Accessories Store	Adult Boutique	Alternative Healer	American Restaurant	Amphitheater	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	R
0	0	0	Seattle	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	Seattle	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	Seattle	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	Seattle	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	Seattle	0	0	0	0	0	0	0	0	0	0	0	0	

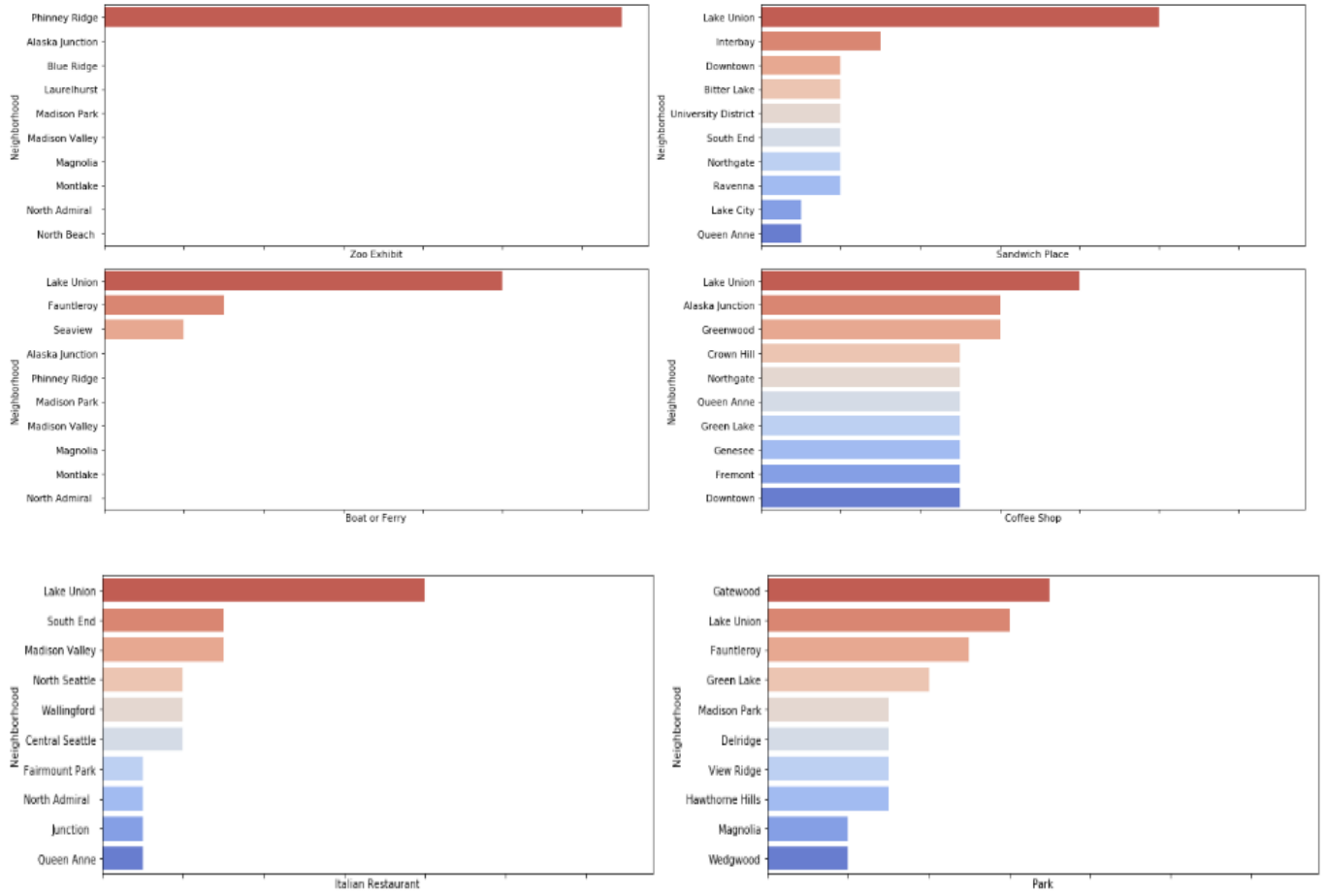
Methodology

The top 10 venue categories can be found by counting their occurrences, the top venue is listed are Zoo Exhibit, sandwich place, boat or ferry, coffee shop and Italian restaurants.

	count	mean	std	min	25%	50%	75%	max
Zoo Exhibit	48.0	0.270833	1.876388	0.0	0.0	0.0	0.0	13.0
Sandwich Place	48.0	0.750000	1.577771	0.0	0.0	0.0	1.0	10.0
Boat or Ferry	48.0	0.312500	1.518065	0.0	0.0	0.0	0.0	10.0
Coffee Shop	48.0	2.895833	2.013144	0.0	1.0	3.0	5.0	8.0
Italian Restaurant	48.0	0.750000	1.328893	0.0	0.0	0.0	1.0	8.0
Park	48.0	1.312500	1.613194	0.0	0.0	1.0	2.0	7.0
Pizza Place	48.0	1.937500	1.743148	0.0	1.0	1.0	3.0	6.0
Bus Stop	48.0	0.354167	1.101055	0.0	0.0	0.0	0.0	6.0
Deli / Bodega	48.0	0.291667	0.944375	0.0	0.0	0.0	0.0	6.0
Beach	48.0	0.375000	0.913842	0.0	0.0	0.0	0.0	5.0

Methodology

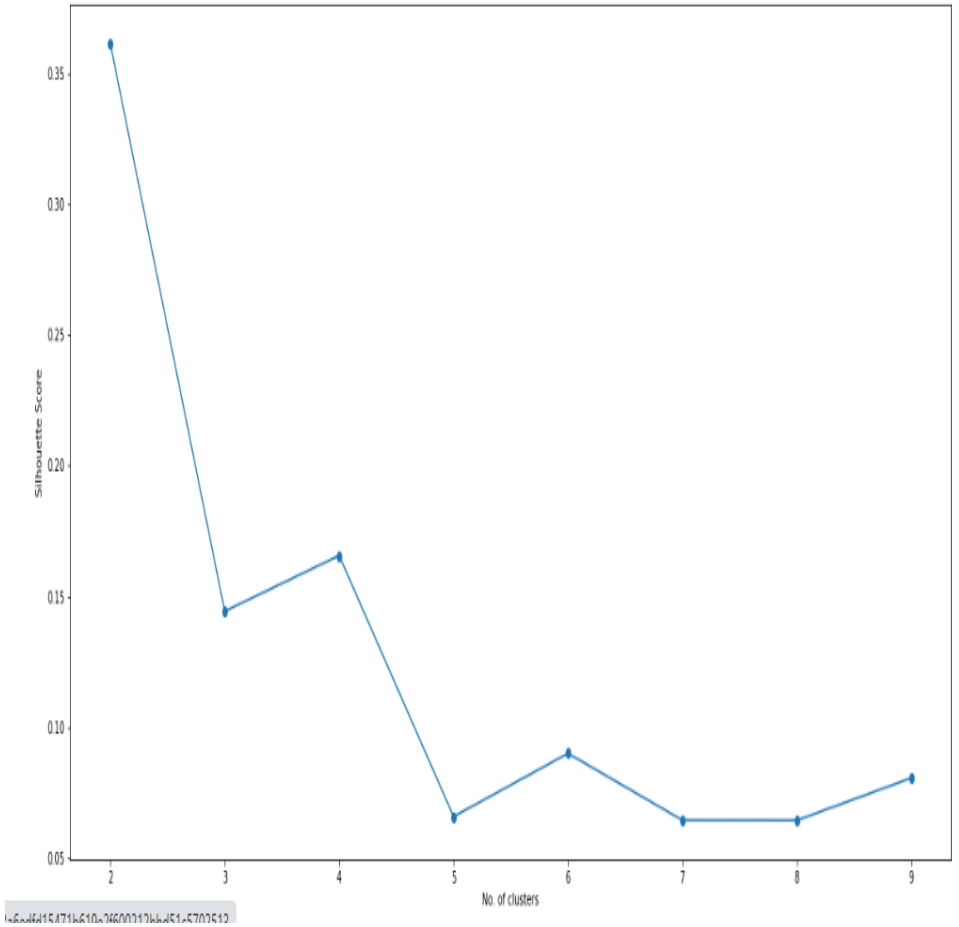
Then the top 10 venues are plotted individually by neighborhoods using seaborn python library for easy visual reading to see possibility on where to be crowded for each neighborhood.



Methodology

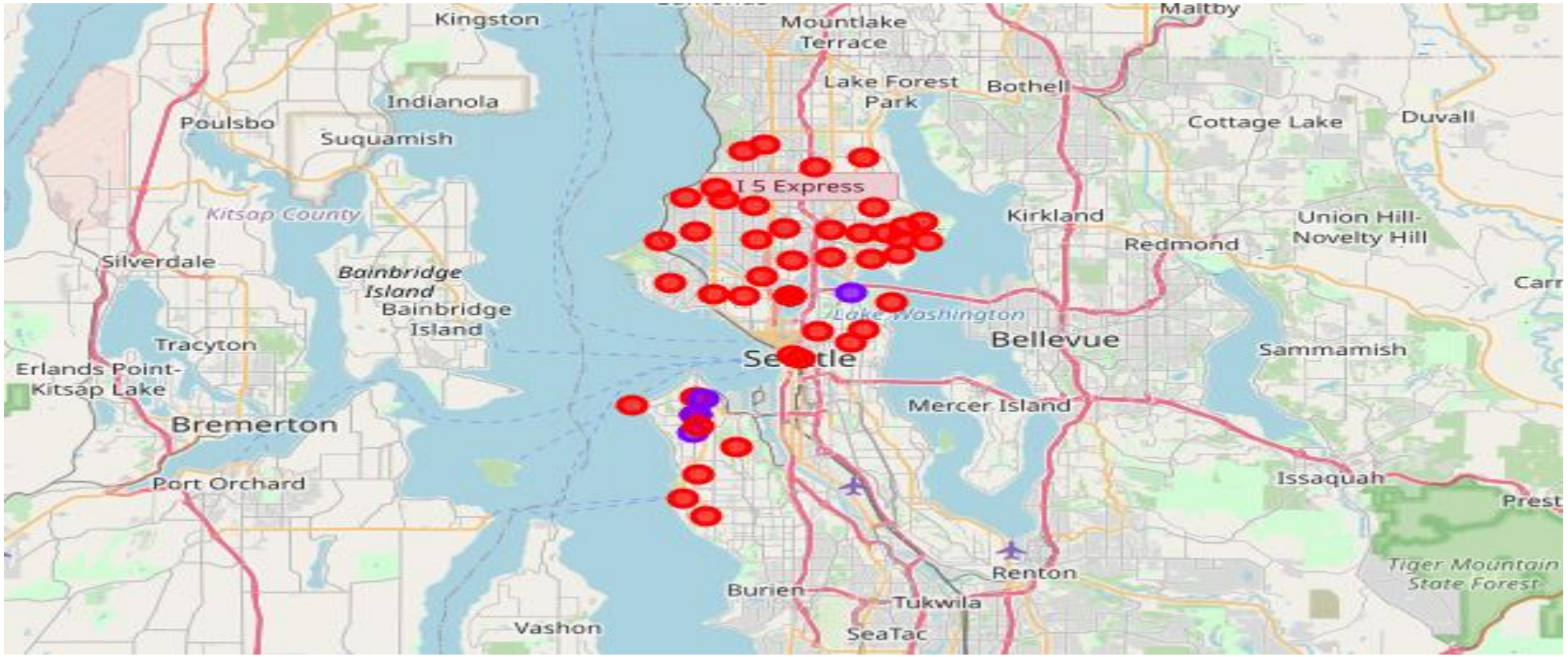
K-Means is an unsupervised ML algorithm which creates cluster of data points based on the similarities among the data point. This algorithm will be used to count neighborhoods for each cluster for variable cluster size.

With initializing the KMeans n to 10, calculate the silhouette_score to find the optimal cluster value,



Methodology

I used folium library again to visualize the Seattle neighborhood with the cluster segmentation included on this,



Results:

There are two clusters, Below are the cluster information on the venue that suppose to be crowded that can be monitored for extra precautions and safety,

Result from cluster analysis – 0

Most 1st common

Coffee Shop	16
Park	6
Pizza Place	6
Boat or Ferry	4
Café	2
Bar	2
Auto Workshop	1

Most 2nd common

Coffee Shop	7
Pizza Place	4
Sandwich Place	4
Park	3
Bank	2
Beach	2
Bar	2

Results:

There are two clusters, Below are the cluster information on the venue that suppose to be crowded that can be monitored for extra precautions and safety,

Result from cluster analysis – 1

Most 1st common

Coffee Shop	2
Pizza Place	1
Trail	1

Most 2nd common

Pizza Place	1
Pub	1
Coffee Shop	1

Discussions:

- Performed cluster analysis with available data information and the coordinates obtain, segmented into two clusters with most common places being coffee shops, restaurants, boat ferry, park, and beach are most crowded places, Not more surprising that both cluster and the most common place look more alike.
- As I mentioned in the data column, the neighborhood information did not contain postal code, assumed that each neighborhood in a borough belong to different postal code, for more detailed and accurate model, the data set can be expanded and verified by obtaining it from specific platform as most current data set.
- This model can be expanded further to other states and cities, that is more for discussion in gathering more accurate data.

Conclusions

During this phase of time, it is crucial to act with more precautions and cautions, the graph shows the consistent upward arch while in the other end thing are getting to normal called “new” normal with precautions. But still people find it difficult to get adjusted to new normal, where bringing more risk prone zones for easy contamination.

The main goal of building this model is to obtain a most crowded places in each neighborhood/borough to set a rules based on the venue to maintain social distancing and keep away to get adjusted to new normal and is our responsibility to bend the curve of COVID-19.

References

- https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle
- https://developers.google.com/maps/documentation/geocoding/
- https://developer.foursquare.com/
- Applied data science capstone lab notebook created by the staffs.