## Clustering and Segmenting New York counties to see Covid-19 Impact :

As a part of IBM's Applied Data Science capstone course on coursera, I worked on a Capstone project where I used Four Square API and machine learning algorithm to fetch the required information and have attempted clustering analysis on it. This report consists of three parts: Business Problem and Data Preparation and cleaning, Methodology, Results and discussion and conclusion.

## Business Problem:

As we know, World has been facing an unprecedented threat of Covid-19. While, it has spread and penetrated rapidly into different parts of the globe, in the United States, New York is one of the worst hit state by COVID-19. New York city is at the centre of the disaster. According to an article, on the webpage, <https://ny.eater.com/2020/12/9/22165800/indoor-dining-nyc-coronavirus-spread-robert-mujica> by Tanay Warerkar and Erika Adams, they mentioned that indoor dining at restaurants and bars are currently the fifth or sixth highest category in the city for the spread of the virus, but outlined that the top categories pushing the city toward a full-blown second wave of the pandemic were not mentioned. In this capestone project, I have attempted to find those top categories. For this research, I have taken the most affected counties in New York with a high number of confirmed cases. Furthermore, by using the four-Square API, we will figure out the top 10 common venues with the most foot traffic.

**Target audience**: After getting this information, a common man will get a glimpse into one of the causes of the rapid spread of the virus.

So, how could we leverage Foursquare location data and machine learning to help us in decision making and point to the evidence that pertains to the spread of coronavirus, so that we can know more about the causes? This is the problem which I am going to discuss by taking New York as an example. Here in this capstone, we will go through each step of this project and address them separately. I will first outline the initial data preparation and describe future steps to start the analysis of venues in New York.

### **Data Preparation:**

For this project we need the following data:

1.United states’ confirmed COVID-19 cases data that contains list of counties (Admin2) of each state with their longitudes and latitudes.  
Data Source : <https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv>

2.Venues in each county of New York

Data Source: Four-square APIs

The project would use Four Square API as its primary data gathering source as it has a database of millions of places. Hence, The Four-Square data is used to extract the venue data centred around the counties of New York. The venue data includes venue name, venue categories and location coordinates. From these two datasets, required venue categories would be identified and clustered using the k-means clustering algorithms.

### **Data Cleaning:**

After examining the United States dataset, let's first import some of the libraries that we’ll be using, and set some parameters to make the output easier to read.

### **Load and Explore the Dataset:**

#### The data source is the list of counties (Admin2) of each state of United States with Province state, latitude (lat) and longitude (long\_) of each county. Here in New York data set link, Admin2 is the County of each state of United states. Now, load data into panda’s data frame to explore the data. Here We are using urllib. Request Python module for fetching URLs. In this module, HTTP is based on requests and responses - the client makes requests and servers send responses. urllib. request mirrors this with a Request object which represents the HTTP request you are making. Once it’s loaded, we’ll want to do some basic cleaning tasks to remove some information we don’t need that will make our data processing slower. Firstly, we have dataset of United states data with county of each state in the United States. Later on, From the final United States dataset, We are retrieve only "New York" province\_state data with all counties having location coordinates and confirmed covid-19 cases. After this, eliminating the rows which has 0 latitudes and longitudes. And then, arrange the confirmed cases in ascending order to know which 10 counties have maximum confirmed cases amongst all. After cleaning the dataset, our final dataset has ready for analysis which contains top 10 Admin2 with most confirmed cases.

**Feature Selection:** Here we are selecting only those features which are required in our analysis that is UID, Admin2(County), Province\_state, Lat (Latitude), Long\_(Longitude), Combined key and total\_confirmed\_cases. And we have dataset from January 2020 to January 2021, but here I am taking only one month data.

Table

Description automatically generated

Chart, bar chart

Description automatically generated

## 

## Methodology:

### Define Foursquare Credentials and Version

Foursquare data is very comprehensive and it powers location data for Apple, Uber etc. For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots around these 10 major counties of New York. The popular spots returned depends on the time when call is made. So, we may get different popular venues depending upon different time of the day. The call returns JSON file and we need to turn that into a data frame. Here, I've chosen 100 popular spots for each major county within a radius of 9 km.

As a first step, we retrieved the data from Four-square APIs. We extract venue information from the centre of New York up to 1000m. The latitude and longitude values are used to fetch venue details. From the JSON file, we get responses of different venues with their names, categories, latitude and longitude. Then we combine our New York dataset with the venues extracted from the Four Square. Next, we'll analyse the data that we created based on the counties with highly confirmed cases of covid-19. We'll identify the frequency of places where there is a lot of foot traffic. Afterwards, we explore the venue categories by plotting the bar graph to see which categories has the most foot traffic and categories which has the least visit by New Yorkers. Lastly, we'll cluster the venue categories based on the available information of each venue. This will allow us to clearly identify which venues could be the reason of the spread and with what characteristics.

## Analysis:

## Categories:

## Chart Description automatically generated

### The bar graph here is not showing all the categories because I took snapshot of only those categories which has high foot traffic. It is clearly seen from the picture doctor’s office, Salon/ Barbershop, Residential buildings/ condo have the maximum frequency of person visit.

### **Most Venues:**

We have different types of venues in each county of New York. We will take a look at the Counties (Admin2) and check which county has the majority of venues with their count.

Table

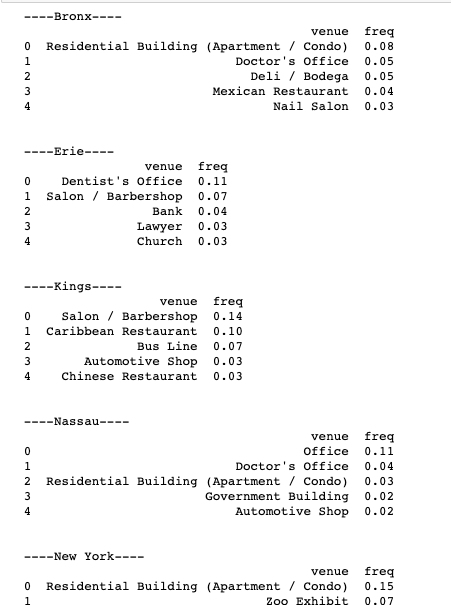
Description automatically generated

From the above results, we can see that New York has most values of active venues in the list and their boroughs like Bronx, kings’ count is near to it. So, from this data, we can predict that New York and their boroughs has the most venues and it could be the centre of disaster.

## One hot Encoding:

## One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. After the one code encoding has done, we have data into numeric form to make machine learning predictions.

Now, we are looking into the top 5 most common venues of each county (Admin2) and the frequency of visitors in each category. This would be a great step to visualize, except restaurants and bars, which other categories are contributing to the spiking of COVID-19 cases. Next, we will merge these common venues with New York data that contains county (Admin2), latitude, longitude to see the top 10 common venues in each county.



### **Cluster Visualization**:

Finally, I have clustered all the venues that fall into their respective county to identify similar venues and the relationship amongst them. I used KMeans clustering and decided to cluster into five groups. However, it is good for us to see the map of the 9 clusters so we can know visually the geographical distribution of all clusters.

Map

Description automatically generated

## Result and Discussion

From the cluster Analysis, we can clearly see that the topmost venue categories of each county where people frequently visit.

Graphical user interface

Description automatically generated

From the cluster 0, we can see Salon/Barbershop, nail shops, college classroom and dentist offices have a lot of foot traffic. Also, the nature of the business conducted at these work cites is something where social distance is quite not possible. Another similar place with social distancing issues is college classroom and Doctor’s office, as the one person alone can be the source of the virus spread to many people. This can also make contact tracing very difficult.

## Cluster1:

## Graphical user interface Description automatically generated

Cluster 1 shows people in New York frequently visit residential buildings/condo, which is the main cause of spreading this virus. The majority of New York City residents live in apartments — in close quarters with neighbours above, below and on either side. And that has added to the challenges of the spread of COVID-19 in the five boroughs and counties of New York. Moreover, close proximity of relatives & friends of the tenants of residential buildings with each other would be another reason. The second most common venue is Deli, we find a lot of foot traffic in these shops too.

## Cluster 2:

## Table Description automatically generated

Cluster 2 again shows salon/Barbershop is the most common venue in the given latitude and longitudes. The second most common venues are restaurants and food in general. As we know, in Salon/barbershop, it's very hard to maintain distance and virus can easily be transmitted from one person to another. Food trucks and restaurants have chefs and as we already explained earlier, if chef is infected, then he can infect all his customers.

## Cluster 3:

## Graphical user interface, table Description automatically generated

Cluster 3 shows different common venues for different counties. In Suffolk, common venues are Farm, food truck, other great outdoors and bus station whereas in Rockland, Doctor's office, church, Salon/Barbershop and Hospitals are common venues. As we know, in Salon/barbershop, it's very hard to maintain distance and virus can easily be transmitted from one person to another. Food trucks have chefs and as we already explained earlier, if chef is infected, then he can infect all his customers.

## Cluster 4:

## Graphical user interface, table Description automatically generated

Cluster 4 shows only one county that is kings where the most common venues are residential buildings/condo, Restaurants, doctor’s office and Food. All these venues have high foot traffic and these are the places where the virus can be easily transmitted because people are in close proximity of each other. It is difficult for one to prevent themselves from getting infected at these crowded places. Because of the preventative measures taken by doctors and employees, the chances of spreading this virus is reduced. But people working over there and patients visiting on regular basis are still prone to getting infected.

**Discussion:**

**Chart, scatter chart

Description automatically generated**

From our analysis, we noticed that some venues are quite common for visitations across all counties of New York particularly residential buildings/condo, barbershops, Nail salons, Deli and restaurants. This analysis gives us the answer of an article mentioned above in the beginning which outlined the missing top four categories while the restaurants and bars fall under the fifth and sixth category of virus spread. Some challenges I have faced while running the Four-square API are that it gives us only venue information from which I can anticipate that these top common venues in each county could be the reason of rapid spread of covid-19. In New York dataset I have access to only confirmed cases of each county instead of each neighbourhood of New York. If I had more information regarding neighbourhoods of New York city then I can construct a better k-Means model. Moreover, if I could have more data regarding neighbourhoods of New York then I could clearly see from the analysis, which areas are more affected in each borough.

**Conclusion**

The purpose of this project was to explore the most affected counties of New York and to get the rough idea of which places this virus has spread. Nothing has fostered the increase of Covid-19 spread more than private family gatherings according to cluster analysis. We can clearly see residential buildings/condo are quite common in each cluster. One of the limitations of this project was that I have confirmed cases of only counties and New York boroughs. However, each borough and county are quite big in size and needs to be broken down to neighbourhoods to get a more accurate representation, but total confirmed covid-19 cases of neighbourhoods of each borough were not available for New York. In this way, it's hard to find out the most affected residential neighbourhoods, but, if I were able to extract contact tracing data then I would be able to make clusters of highly affected residential areas.

After merging data from the Four-Square API, I have the performed EDA to reveal a few interesting observations and I have also performed K means clustering on the data set based on different counties confirmed cases. It revealed us possible hotspots for the spread of virus. After attempting the one-hot encoding, I can see top 5 venues of each county with frequency of visitors. Furthermore, the clustering exercises highlighted the group of optimal areas and gives us the glimpse into reason of spreading this virus rapidly.