COVID 19 SPREAD INSIGHTS FOR THE US

IBM Applied Data Science Certification Capstone Project

ABSTRACT

This is the initial report of the COVID 19 Insights Capstone Project.

Rajesh Ramachandran

Coursera IBM Applied Data Science Capstone

s

CONTENTS

INTRODUCTION

The world has never experienced such a widespread pandemic like COVID-19. The uncertainty over this epidemic has marred nations. Fortunately, we live in an era, where unimaginable data is being generated every second. To tackle COVID -19 and make intelligent decisions, organizations are heavily relying data that is being generated and assembled by various sources in the world. More and more insights driven by machine learning algorithms and fast data processing has enabled to derive strategies and action plans.

A platform to provide COVID 19 case spread at various levels (County, State, National) insights and visualizations help frontline working strategies to be streamlined. Targeted programmes like increase in testing centers, can be deployed to areas affected worse and with minimal healthcare facilities.

BUSINESS PROBLEM

The attempt of this Capstone Project is to find and analyze COVID 19 data available in the public forum for the United States. Using Data Science Methodologies, the outcome of this analysis we will try to answer these business questions.

1. Which are the worst and least affected areas in the US?
2. Can we classify Counties of any US State into High, Medium, Low risk based on their health care infrastructure?

TARGET AUDIENCE FOR THIS PROJECT

Mainly the project is targeted at helping Health Care Organizations and teams tackling the COVID19 pandemic from a planning perspective. It includes, Hospitals, State and National Healthcare officials and any organizations involved in COVID 19 frontline work. It helps them to make data-driven decisions.

In addition to that this project is also aimed anyone who is directly or indirectly affected by COVID 19 who would like to get a perspective of the COVID 19 spread.

DATA SOURCES

The problem presented above would need the following data from available sources

1. Daily COVID 19 Cases reported data in the US. **The John Hopkins University Center for Systems Science and Engineering Data** repository will be used as source for this purpose. The GitHub page for JHU CSSE has multiple datasets in CSV format with active reported cases of COVID 19 throughout the world. The datasets are available here (https://github.com/CSSEGISandData/COVID-19).
2. For the classification problem, we will need data from 2 sources
    a. To get a list of cities in each county we will be using a publicly available GitHub US cities and Counties database in CSV format. The dataset is available here (https://github.com/grammakov/USA-cities-and-states/blob/master/us_cities_states_counties.csv)
    b. To get Hospital venue information for each city, we would be accessing the Foursquare Database using the Foursquare Developer API. Foursquare has one of largest database of 105+ million places and is used by over 125,000 users worldwide. More info on Foursquare APIs can be found here (https://developer.foursquare.com/docs/places-api/)

1. JHU CSSE COVID-19 DATASET

For our project we will be using the CSSE COVID 19 Daily Report, which is uploaded daily in the CSSE GitHub page.

Daily reports (csse_covid_19_daily_reports))

This table contains an aggregation of each USA State level data.

FILE NAMING CONVENTION

MM-DD-YYYY.csv in UTC.

FIELD DESCRIPTION

1. **FIPS**: US only. Federal Information Processing Standards code that uniquely identifies counties within the USA.
2. **Admin2**: County name. US only.
3. **Province_State**: Province, state or dependency name.
4. **Country_Region**: Country, region or sovereignty name. The names of locations included on the Website correspond with the official designations used by the U.S. Department of State.
5. **Last Update**: MM/DD/YYYY HH:mm:ss (24 hour format, in UTC).
6. **Lat and Long** : Dot locations on the dashboard. All points (except for Australia) shown on the map are based on geographic centroids, and are not representative of a specific address, building or any location at a spatial scale finer than a province/state. Australian dots are located at the centroid of the largest city in each state.
7. **Confirmed**: Confirmed cases include presumptive positive cases and probable cases, in accordance with CDC guidelines as of April 14.
8. **Deaths**: Death totals in the US include confirmed and probable, in accordance with CDC guidelines as of April 14.
9. **Recovered**: Recovered cases outside China are estimates based on local media reports, and state and local reporting when available, and therefore may be substantially lower than the true number. US state-level recovered cases are from COVID Tracking Project.
10. **Active**: Active cases = total confirmed - total recovered - total deaths.
11. **Incidence_Rate**: confirmed cases per 100,000 persons.
12. **Case-Fatality Ratio (%)**: = number recorded deaths / number confirmed cases

2. US CITIES AND COUNTIES

This is a GitHib respository of US Cities, Counties and States. it contains a full list of US cities, states and counties (60k+ entries) formatted. The file is available in pipe delimited format here

FILE STRUCTURE

City|State short name|State full name|County|City Alias Mixed Case

EXAMPLE

Bronx|NY|New York|BRONX|University Heights

3. FOURSQUARE DATABASE

The Foursquare Database is used in the is projects to get the number of hospitals per county. Since County is a big geographical entity we will be searching the Foursquare API for hospital venues at city level and group them at county level later.

The Foursquare Database is accessed via the Foursquare Developer API which required credentials. The credentials have been previously setup and kept in secure python file, which is used by this project.

We will be calling the API iteratively till we get all hospital venues from the API within the given distance. The maximum venues the API can fetch is 100, so we will fetch all venues by iteratively calling this API and increasing the offset each time.

We will be using the Foursquare Venue Search API. Credentials required to access the Foursquare API are

- CLIENT_ID
- CLIENT_SECRET
- VERSION

Parameters Required for the Venue Search API

- **ll** = Latitude and longitude of the search location. This will be the Lat and Longitude of the city we wish to search for.
- **radius** = currently we will set this as 20 miles (~32000 meters). Assumption here is that each city is at the max 40 miles wide.
- **categoryId** = A comma separated list of categories to limit results to. In our case this will be list of categories for hospitals is
    - **Hospital** - 4bf58dd8d48988d196941735
    - **Hospital Ward** - 58daa1558bbb0b01f18ec1f7
    - **Medical Lab** - 4f4531b14b9074f6e4fb0103
    - **Urgent Care Center** - 56aa371be4b08b9a8d573526
    - **Doctor's Office** - 4bf58dd8d48988d177941735
    - **Emergency Room** - 4bf58dd8d48988d194941735
    - **Medical Center** - 4bf58dd8d48988d104941735

STRUCTURE OF THE API

The API request URL would look like this

https://api.foursquare.com/v2/venues/search?client_id=XXXX&client_secret=XXXX&v=XXXX&ll=XX,XX&radius=32000&categoryId=LIST_OF_CATEGORIES

METHODOLOGY

DATA PREPARATION

COVID-19 DATASET

- Apply filter to the COVID-19 data set to filter out all countries except the US.
- Rename a few columns for readability
- Remove rows with County as blanks or NaN
- For the Lat, Lng columns we will remove the rows which contain NaN values
- For the other metric columns, we will replace NaN with 0s
- Remove duplicates
- Reset Index

The resulting Data Frame would look like this

	FIPS	County	State	Country	Last_Update	Lat	Lng	Confirmed	Deaths	Recovered	Active	Combined_Key	Incidence_Rate	Case-Fatality_Ratio
0	45001.0	Abbeville	South Carolina	US	2020-07-17 04:34:50	34.223334	-82.461707	170	2	0	168.0	Abbeville, South Carolina, US	693.113711	1.176471
1	22001.0	Acadia	Louisiana	US	2020-07-17 04:34:50	30.295065	-92.414197	1574	48	0	1526.0	Acadia, Louisiana, US	2536.868402	3.049555
2	51001.0	Accomack	Virginia	US	2020-07-17 04:34:50	37.767072	-75.632346	1045	14	0	1031.0	Accomack, Virginia, US	3233.692289	1.339713
3	16001.0	Ada	Idaho	US	2020-07-17 04:34:50	43.452658	-116.241552	5132	30	0	5102.0	Ada, Idaho, US	1065.643383	0.584567
4	19001.0	Adair	Iowa	US	2020-07-17 04:34:50	41.330756	-94.471059	19	0	0	19.0	Adair, Iowa, US	265.659955	0.000000

US CITIES AND COUNTIES DATASET

- Apply filter to the US City Counties data set to filter out all States except the CA.
- Rename a few columns for readability
- Remove rows with City/County as blanks or NaN
- For the Lat, Lng columns we will remove the rows which contain NaN values
- Remove duplicates
- Reset Index

Later we will use the GEOPY python library to add coordinate for each city. This will help us browse for venues in each city in the next section where we will use the Four Square API. After adding the GEOPY coordinates the Data Frame looks like this.

	CITY	STATE	STATE_FULL	COUNTY	LAT	LNG
0	Los Angeles	CA	California	LOS ANGELES	34.053691	-118.242767
1	West Hollywood	CA	California	LOS ANGELES	34.092301	-118.369289
2	Playa Vista	CA	California	LOS ANGELES	33.976010	-118.418165
3	LA	CA	California	LOS ANGELES	36.701463	-118.755997
4	Bell Gardens	CA	California	LOS ANGELES	33.969456	-118.150395

GETTING HOSPITAL NUMBERS FROM FOURSQUARE

The Foursquare Database is used in the is projects to get the number of hospitals per county. Since County is a big geographical entity, we will be searching the Foursquare API for hospital venues at city level and group them at county level. We will be running the Four Square API in loop for all the rows City-County dataset which was illustrated previously

As discussed before we will be only searching for the following categories.

- **Hospital** - 4bf58dd8d48988d196941735
- **Hospital Ward** - 58daa1558bbb0b01f18ec1f7
- **Medical Lab** - 4f4531b14b9074f6e4fb0103
- **Urgent Care Center** - 56aa371be4b08b9a8d573526
- **Doctor's Office** - 4bf58dd8d48988d177941735
- **Emergency Room** - 4bf58dd8d48988d194941735
- **Medical Center** - 4bf58dd8d48988d104941735

After running the Four Square API for each coordinate in the City-County dataset we get a list of all hospitals in each coordinate (city). We then group the entire list at a County level which will get us the number of hospitals in each County in CA. In esscence we will be doing these

- Remove all duplicate coordinates generated
- Reset Index
- Group at State/County level for sum of hospitals, in each county

Next, we will merge the County grouped dataset to the main COVID 19 dataset (filtered on CA) to add the hospital count for each county. This will be our final CPVID 19 California dataset that we will use to gain insights thru classification.

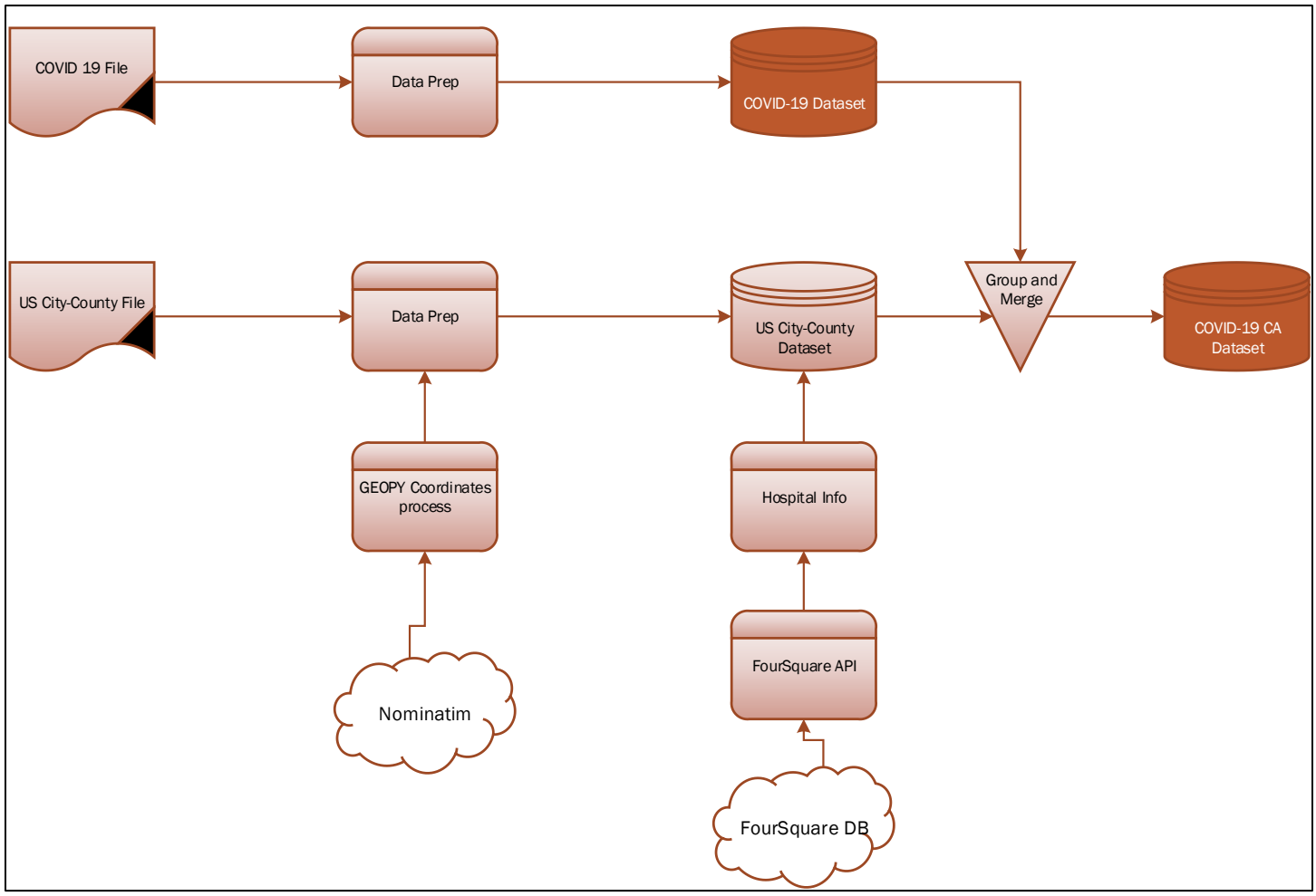
Here we are doing

- Convert County column to uppercase
- Join the filtered COVID dataset to County Dataset
- Reset index
- Add GEO_ID column for visualization purposes

This is how it will look like

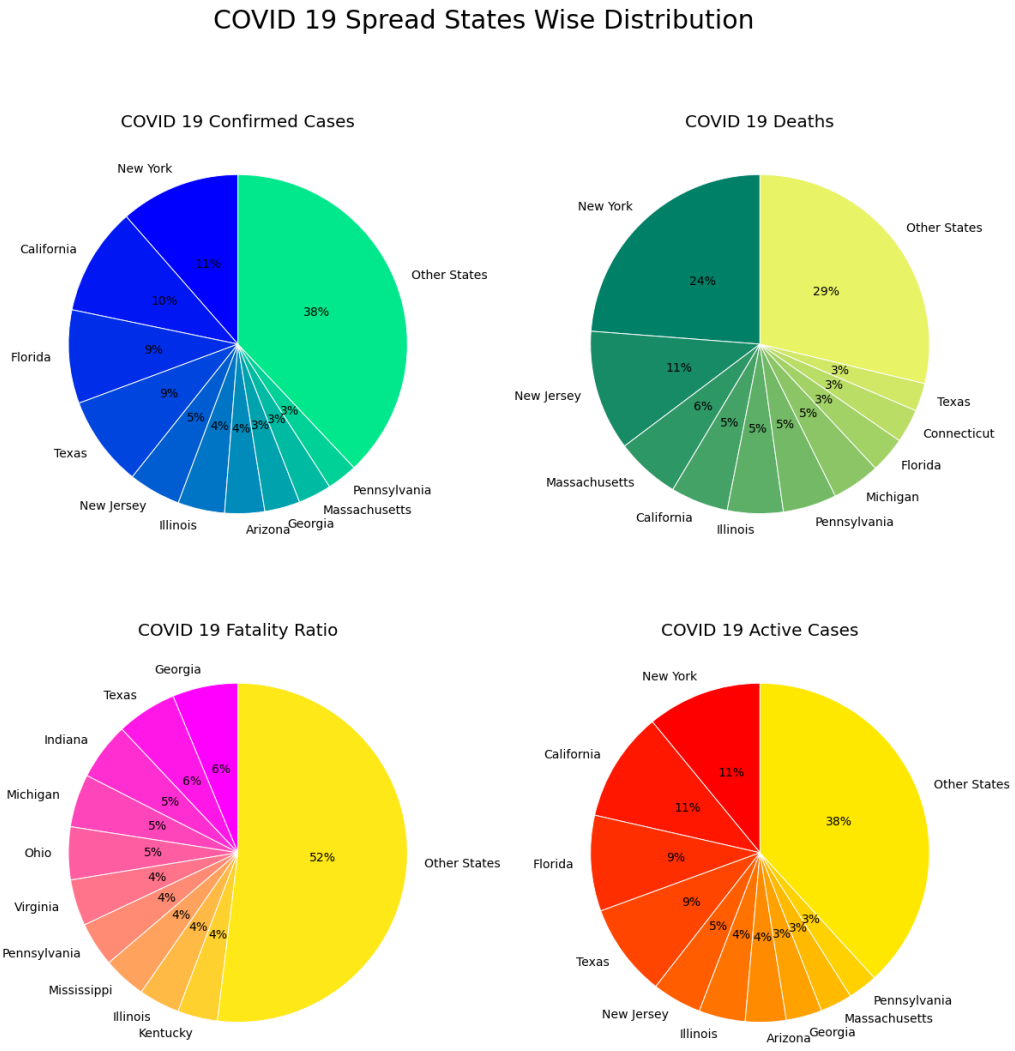
	County	FIPS	State	Country	Last_Update	Lat	Lng	Confirmed	Deaths	Recovered	Active	Combined_Key	Incidence_Rate	Case-Fatality_Ratio	HOSPITAL_COUNT	GEO_ID
0	ALAMEDA	6001.0	California	US	2020-07-17 04:34:50	37.646294	-121.892927	8627	154	0	8473.0	Alameda, California, US	516.176049	1.785093	79	0500000US06001
1	ALPINE	6003.0	California	US	2020-07-17 04:34:50	38.596786	-119.822359	2	0	0	2.0	Alpine, California, US	177.147919	0.000000	36	0500000US06003
2	AMADOR	6005.0	California	US	2020-07-17 04:34:50	38.445831	-120.656960	49	0	0	49.0	Amador, California, US	123.264238	0.000000	106	0500000US06005
3	BUTTE	6007.0	California	US	2020-07-17 04:34:50	39.667278	-121.600525	487	4	0	483.0	Butte, California, US	222.185724	0.821355	71	0500000US06007
4	CALAVERAS	6009.0	California	US	2020-07-17 04:34:50	38.205371	-120.552913	75	0	0	75.0	Calaveras, California, US	163.380895	0.000000	75	0500000US06009

The entire data preparation process has been illustrated in the flow chart below.

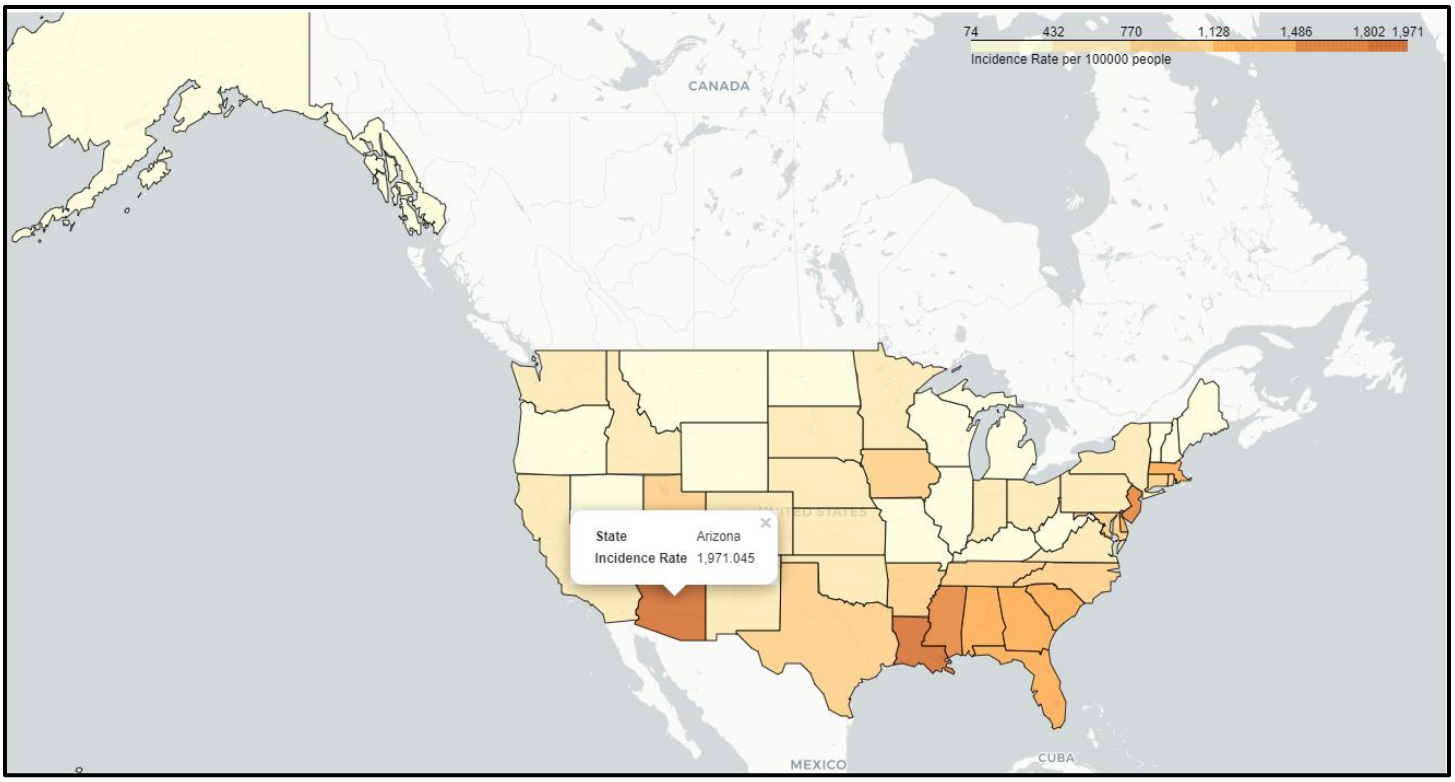


DATA ANALYSIS

With all the datasets cleaned up and prepped, we can now start analyzing the data. We first look at the COVID 19 US Stats and see what the spread across top 10 states for the different case metrics is. We will be analyzing Confirmed Case, Deaths, Active Cases, Case-Fatality Ratio and see what the top grossing states. A pie plot is ideal to depict this

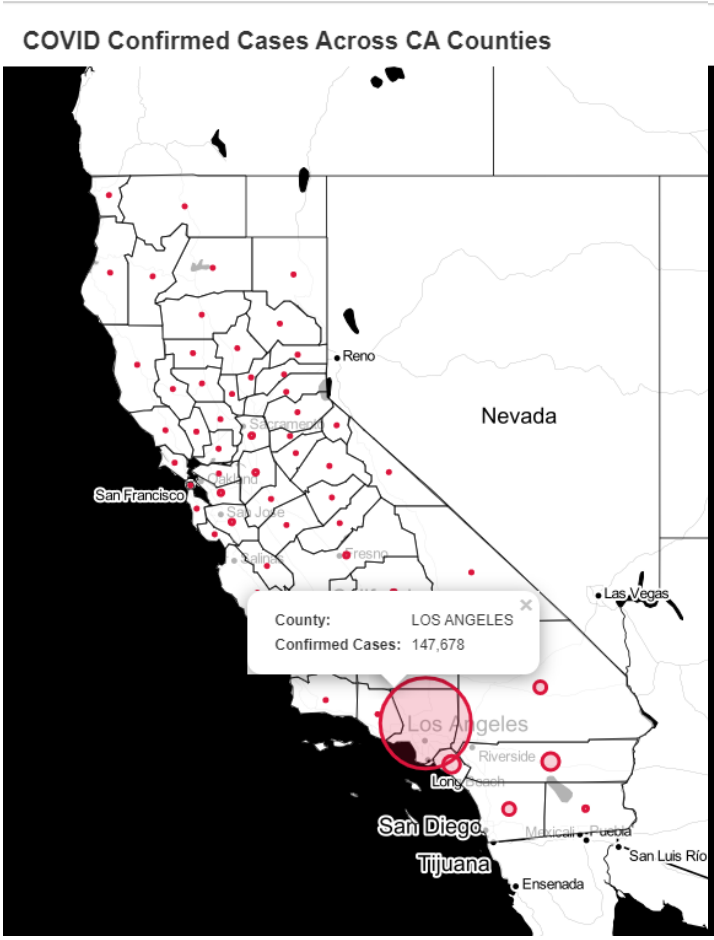


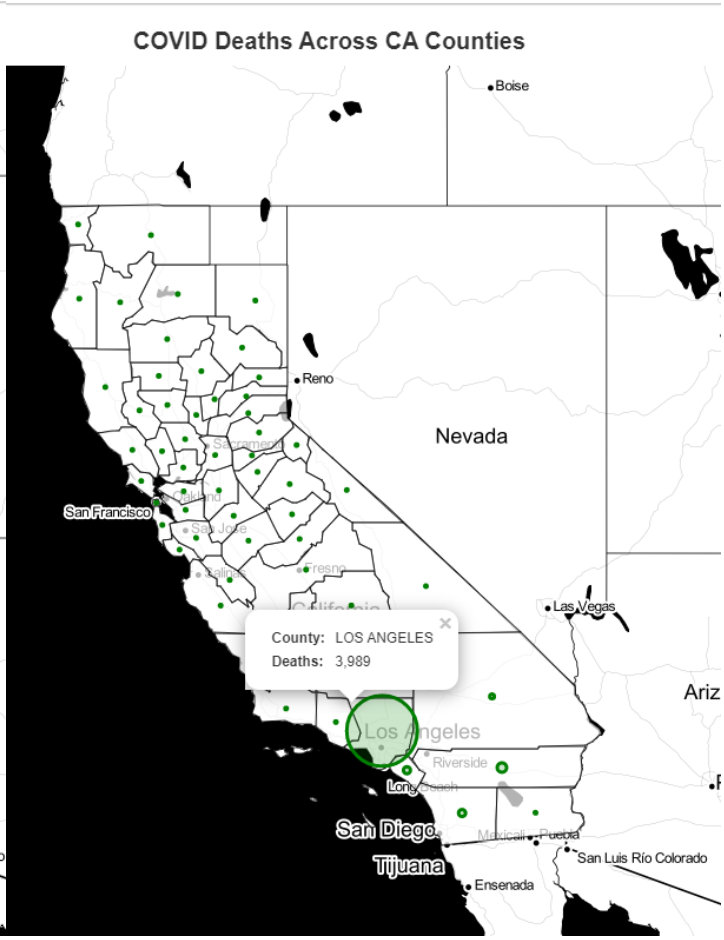
We can derive from thee plots that, state New York tops in Confirmed, Deaths and Active Cases. Georgia tops in Fatality ratio. Other states follow, California and Florida are just right behind New York.



For Incidence rates, we have shown to plot it using a Choropleth map, Arizona, Louisiana and New Jersey tops the incidence rates.

Next, we analyze the county data for California, we will be only dealing with Confirmed Cases and Deaths. Lets plot the data in bubble plot map.



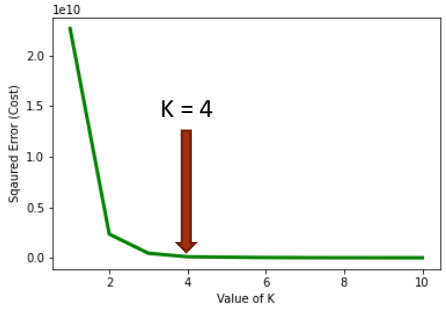


Los Angeles County tops in Confirmed Cases and Deaths with Riverside, San Bernardino and San Diego following.

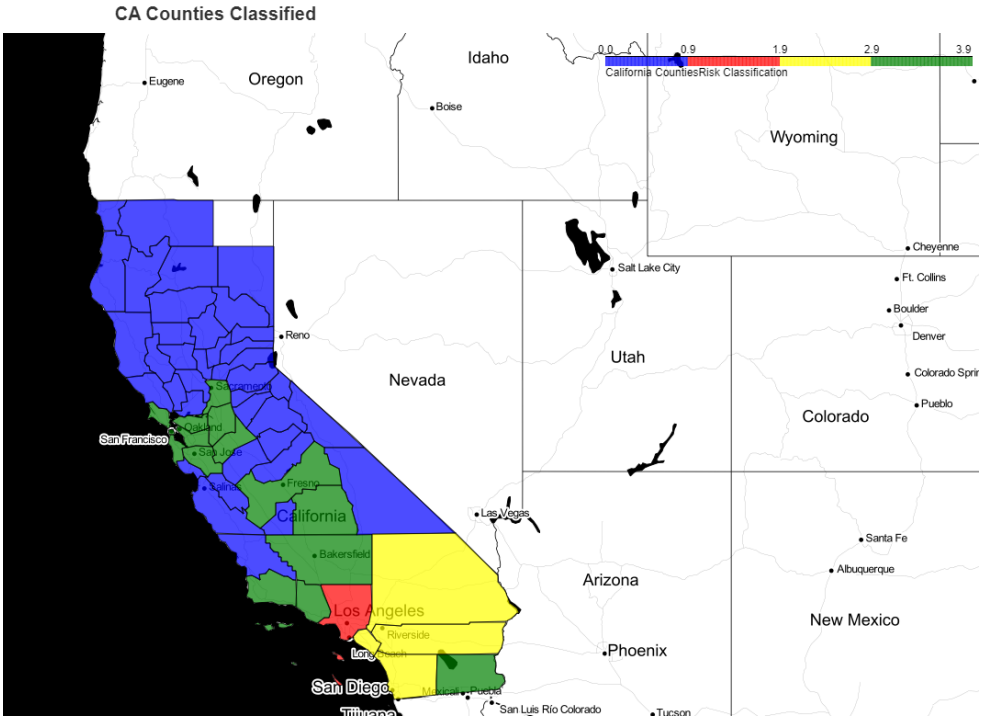
RISK LEVEL ASSESSMENT OF CA COUNTIES

We are going to mark each county based on their risk level. We are going to group each county in groups based on their Confirmed Cases, Deaths and number of Hospitals. For this we will be using the K Means clustering algorithm. To find the optimal K Value for the model, we will use the elbow method.

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes MSE for all clusters. These MSE values are plotted against the K values and the resultant line plot is obtained. If the line plot resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. We plotted out the elbow chart and concluded that K=4 gives the optimal solution.

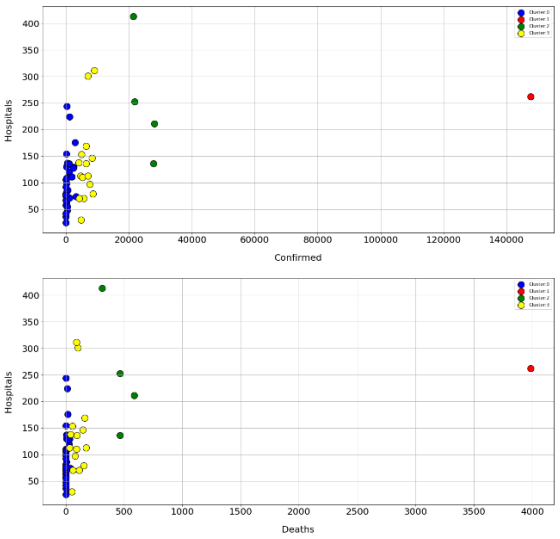


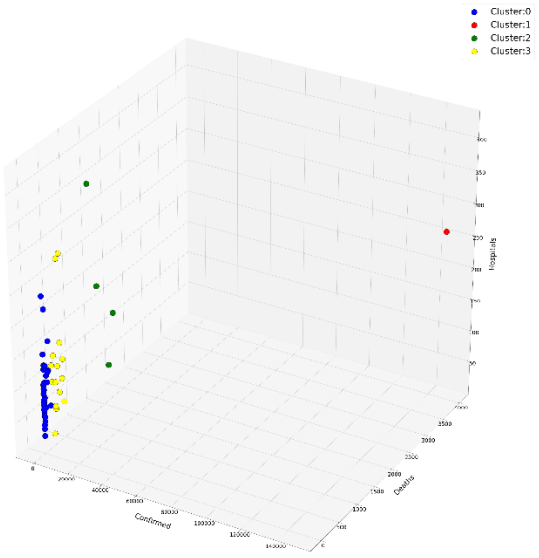
We classified the CA COVID dataset into 4 clusters and assigned them to different Risk Levels. The same was plotted in a map.



RESULTS DISCUSSION

For discussion we initially plot the clusters in scatter plot and view the data dispersion.





As you can see most of the Counties fall under **Cluster 0 (Blue)**. These are counties with the least number of Deaths and Confirmed Cases with a relatively large number of hospitals. So we can categorize them into **Low Risk Counties**. **Cluster 1 (Red)** seems to be an anomaly because of the extremely large count of Deaths and Confirmed case in the Los Angeles County. Due to its relatively average number of Hospitals we will Los Angeles under **High Risk County**. **Cluster 3 (Yellow)** Counties are little bit higher than **Cluster 0 (Blue)** in terms of the COVID Cases with relatively the same spread of Hospitals as the Cluster 0. We will put them under **Below Average Risk** Counties. **Cluster 2 (Green)** is relatively thinly spread with only 4 Counties with extremely large COVID Cases reported and having large number of hospitals. Even though it has large number hospitals the cases numbers are high, hence we will classify them into **Above Average Risk** Counties.

The below table represents the clusters classifications based on the discussion above

Cluster	Classification
0	Low Risk
1	High Risk
2	Above Average
3	Below Average

CONCLUSION

This project was initially started at the aim of answering these questions.

1. How is the COVID 19 Spread in the US? Which are the worst affected areas?
2. Can we analyze and assess risk levels of various counties using the available data?

The projects successfully attempted and provided handsome insights on the COVID spread in form of visualizations and data presentation. We have also tried to asses a risk level of each counties in California using a methodology and strategy devised. We used clustering to group counties in various areas of risk.

With the current data we have tried to project the above-mentioned insights. Further exploring in this realm will be to use patient health data, COVID testing data, population data to provide a better clustering to assess risk levels.