Tracking the coronavirus in my neighborhood

Contents

Introduction:

Background

It is almost nine months in this COVID-19 pandemic, and the world is still looking for some good news from our medical experts and vaccine research. Everyone in the world is affected by the devastating effect of COVID-19. And there are many datasets made available to capture trends of this virus.

I was too interested in analyzing the COVID-19 data in my home state of California and surrounding my neighborhood. I was able to find good publicly available data sources on LA times GitHub directory, which provides daily updates on confirmed cases in each county. I took data that was updated with longitude and latitude coordinates that can be utilized to populate on the California map. There are many data available but difficult to form a mental image of daily trends and neighborhood comparison, along with an increase in cases with time since March 2020.

A brief history of COVID-19 in California: On August 1st, 2020, California became the first state in the U.S. to surpass 500,000 coronavirus infections, capping a month in which confirmed cases rose by nearly two million nationwide. It started on March 4th, 2020, where California reported first death, and on August 8th, 2020, death level reached to 10,293.

Problem Description

Now let me explain the context of this Capstone project through data and following through county wise cases. Say you live in Orange County, which is south part of California and find out how pandemic effect spread through your county and compare to other counties. Also, if you would like to find out how safe your city is compared to the neighboring city, such information is complicated to visualize from the data set. County wise results only provide total cases in each city, but time series analysis is not available.

Wouldn't it be great if you able to determine the spread of the virus in your area compared to other cities and make the wise decision to shop, travel, and move around safely? Also, would you like to find single day cases in the neighborhood compare to another city or county? This project will do it for you.

Objective

This report aims to study and analyze the neighborhoods of California, its county, and cities and compare them into similar county or city and, to analyze those clusters to gather meaningful information. That information on COVID-19 via this project can be used to find out the neighborhood's comparison or single-day hike in cases and pinpoint the least infected area in the state of California.

Target Audience

This information provided by this report would be useful for people who are interested in finding California COVID-19 cases in their city and finding the spread of the virus from the beginning. For example, if the user would like to find a virus spread after summer vacation or 4th of July holidays, it can give a visual impression of each date with single day cases on a single graph, which may be challenging to find and required more data research.

Data Description:

To consider the objective stated above, we can list the data sources used for the analysis.

a) California Coronavirus LA Times Data: The following link page used to pull out the necessary data for this project, more data sets are available at https://github.com/datadesk/california-coronavirus-data. The information obtained, i.e., date, county, place, and confirmed cases with latitude and longitude. The data was transformed into a pandas data frame for further analysis and use Plotly and folium for visualizing data on the map.

b) Coordinate data for each Neighborhood in California: The following CSV file gave us the geographical coordinates of each place in the county: https://raw.githubusercontent.com/datadesk/california-coronavirus-data/master/latimes-place-totals.csv

Methodology:

Importing / Installing library for our project

```
import datetime as dt
import matplotlib.pyplot as plt
from matplotlib import style
import pandas as pd
import altair as alt
import folium as folium
```

Extracting data into a Pandas data frame for California and Irvine

The same data set will also fetch the coordinate data for all the neighborhoods in California using the CSV file and put it into a data frame.

	date	county	fips	place	confirmed_cases	note	x	y
0	2020-08-09	Yolo	113.0	Davis	179	NaN	-121.738056	38.553889
1	2020-08-09	Yolo	113.0	Unincorporated	209	NaN	NaN	NaN
2	2020-08-09	Yolo	113.0	West Sacramento	486	NaN	-121.530278	38.580556
3	2020-08-09	Yolo	113.0	Winters	73	NaN	-121.970833	38.525000
4	2020-08-09	Yolo	113.0	Woodland	774	NaN	-121.773333	38.678611

We can go through each county and put it into a data frame.

	date	county	fips	place	confirmed_cases	note	x	y
97688	2020-03-27	Orange	59.0	Irvine	33	NaN	-117.8436	33.686502
97384	2020-03-28	Orange	59.0	Irvine	36	NaN	-117.8436	33.686502
97072	2020-03-29	Orange	59.0	Irvine	38	NaN	-117.8436	33.686502
96722	2020-03-30	Orange	59.0	Irvine	43	NaN	-117.8436	33.686502
96341	2020-03-31	Orange	59.0	Irvine	50	NaN	-117.8436	33.686502

Data Cleanup

Data provided by the LA Times are missing some information and filled with 'NaN.' We need to clean up this data before we analyze them.

```
ca_county.query("y == 'NaN'")
```

	date	county	fips	place	confirmed_cases	note	x	y
1	2020-08-09	Yolo	113.0	Unincorporated	209	NaN	NaN	NaN
5	2020-08-08	Alameda	1.0	Address unknown	86	NaN	NaN	NaN
9	2020-08-08	Alameda	1.0	Castro Valley	338	NaN	NaN	NaN
14	2020-08-08	Alameda	1.0	Homeless	117	NaN	NaN	NaN
20	2020-08-08	Alameda	1.0	Remainder of county	19	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...
98227	2020-03-24	Los Angeles	37.0	Pico	1	NaN	NaN	NaN
98869	2020-03-18	Los Angeles	37.0	Smaller Los Angeles neighborhoods	62	NaN	NaN	NaN
98916	2020-03-17	Los Angeles	37.0	Smaller Los Angeles neighborhoods	27	NaN	NaN	NaN
98950	2020-03-16	Los Angeles	37.0	Santa Clarita and Stevenson Ranch	3	NaN	NaN	NaN
98954	2020-03-16	Los Angeles	37.0	Smaller Los Angeles neighborhoods	11	NaN	NaN	NaN

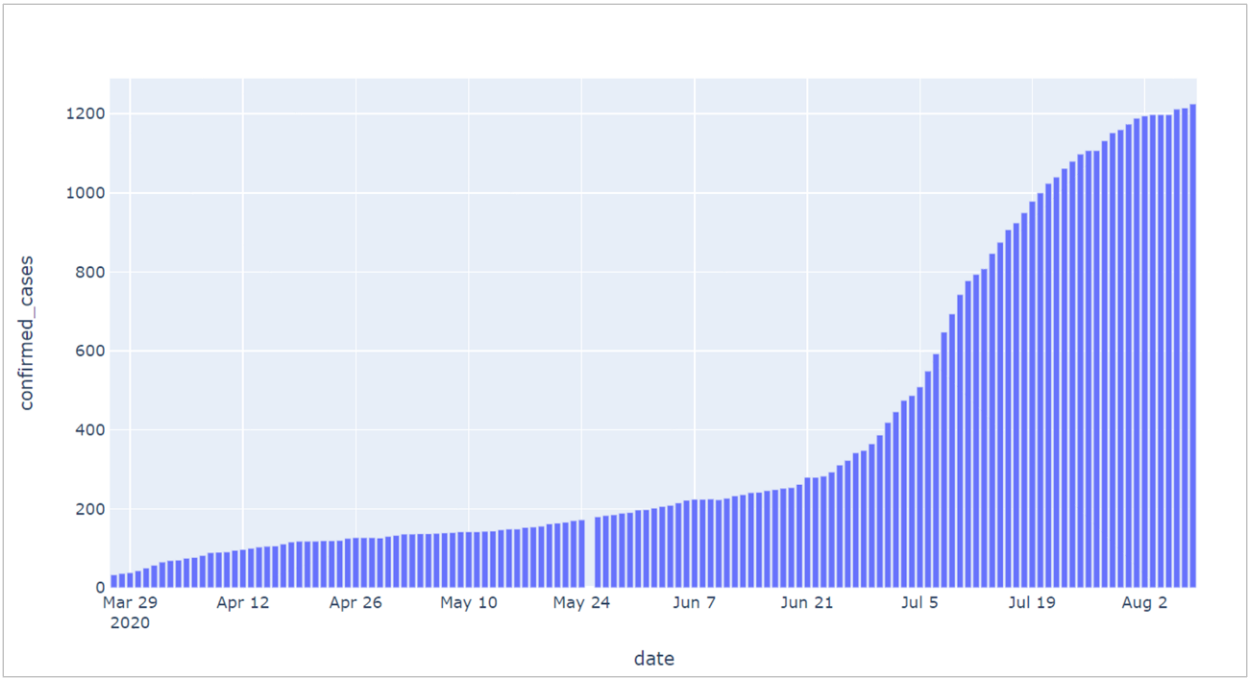
3473 rows × 8 columns

Cleaned data after removing missing information

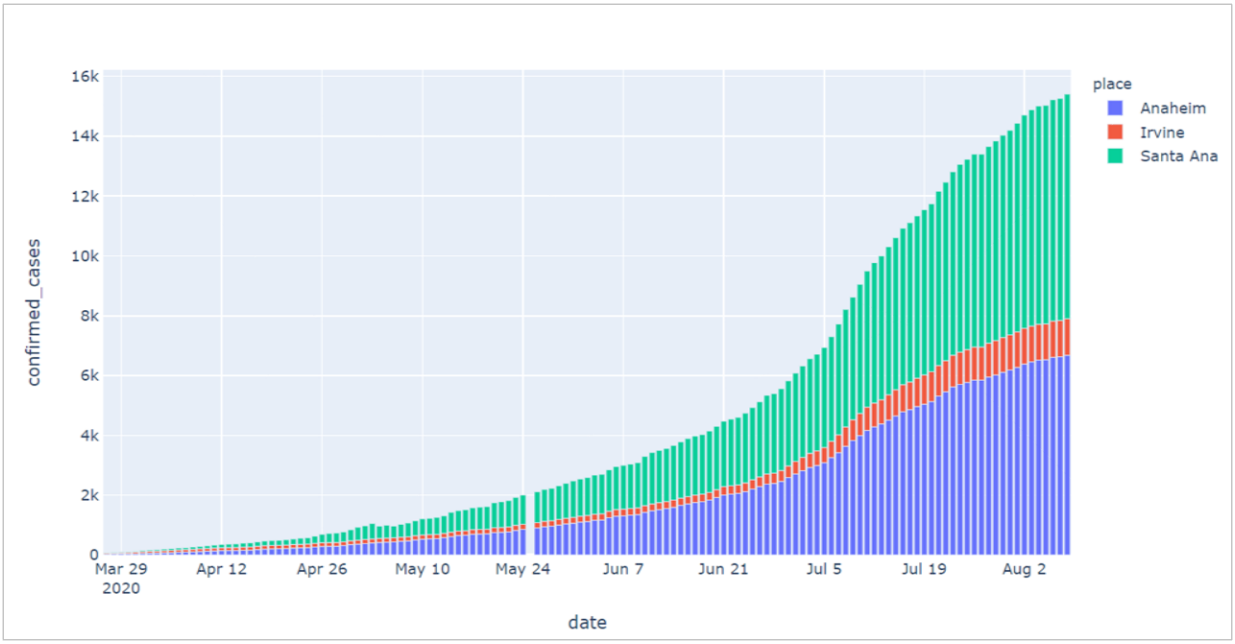
```
ca_county = ca_county.query("confirmed_cases != 'NaN' & x < 0 & x != 'NaN' & date.notnull()", engine='python')
ca_county.head()
```

	date	county	fips	place	confirmed_cases	note	x	y
0	2020-08-09	Yolo	113.0	Davis	179	NaN	-121.738056	38.553889
2	2020-08-09	Yolo	113.0	West Sacramento	486	NaN	-121.530278	38.580556
3	2020-08-09	Yolo	113.0	Winters	73	NaN	-121.970833	38.525000
4	2020-08-09	Yolo	113.0	Woodland	774	NaN	-121.773333	38.678611
6	2020-08-08	Alameda	1.0	Alameda	188	NaN	-122.274444	37.756111

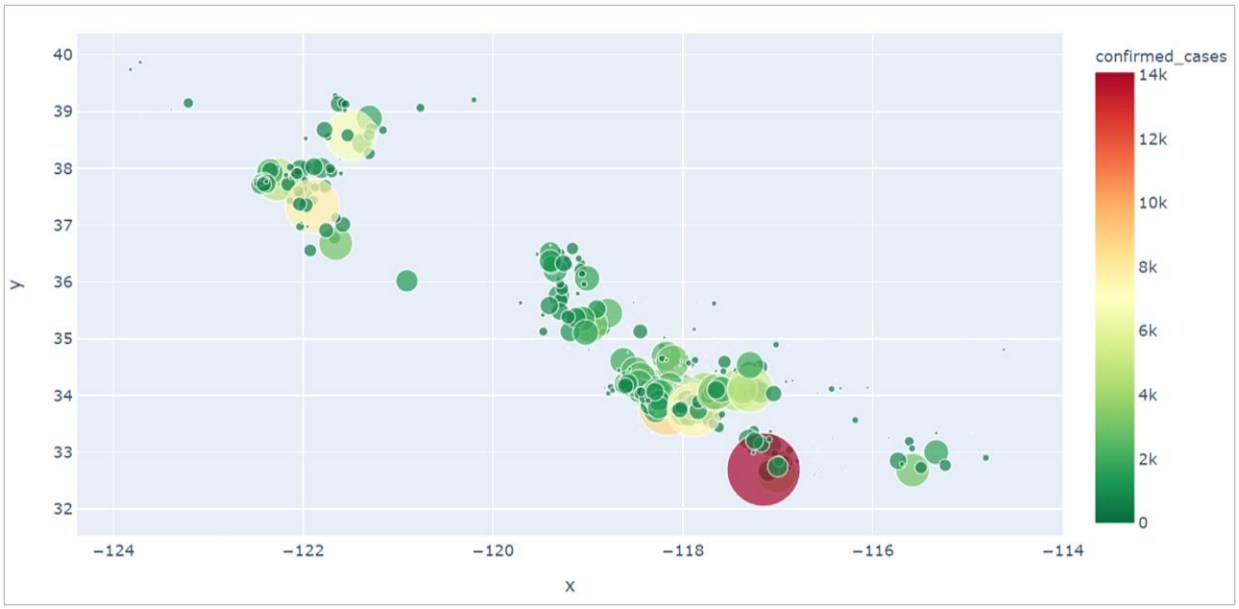
Exploring Irvine coronavirus data with bar chart using Plotly



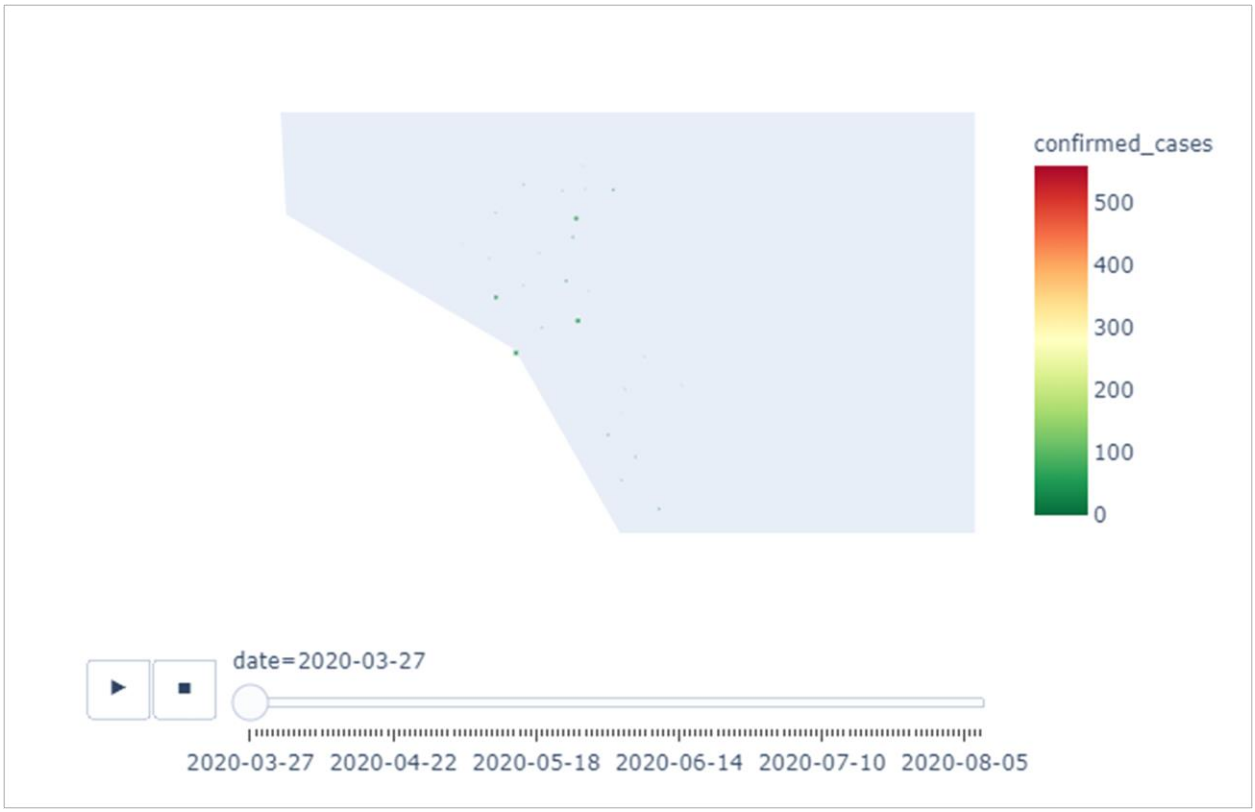
We are exploring Orange County coronavirus data with colors to represent different neighborhoods in our stacked chart.



Exploring California coronavirus data with colors to represent different neighborhoods in the scatter chart overlay on California map. San Diego is the highest number of confirmed cases.



Animated scatter chart of coronavirus data. The '*scatter_geo*' method puts the data on a California map with colors to represent different neighborhoods. You can play the data to see the animated effect with time-lapse.



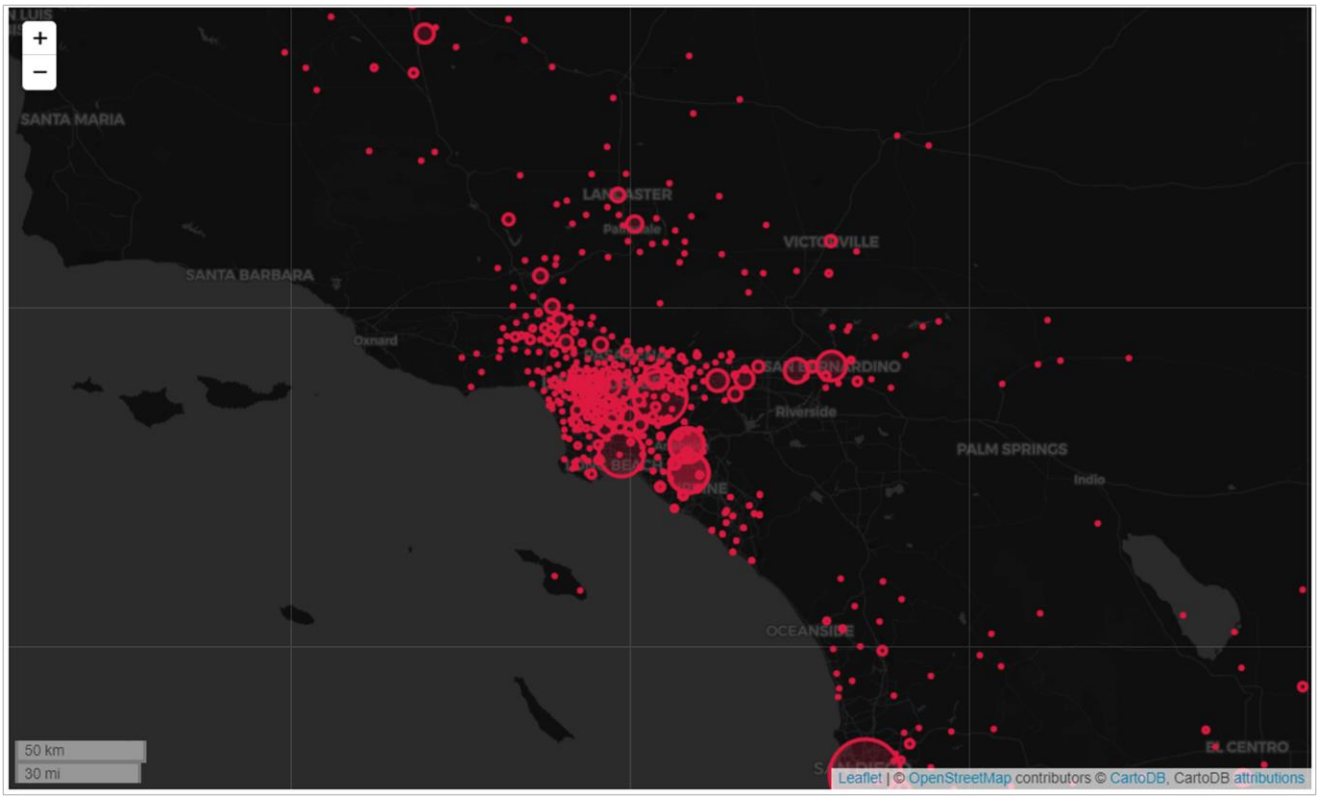
Generating a map of California with center on Irvine and plotting the Neighborhood data on it

We first filter the data with eliminate 'NaN' and generate different data frames like ca_covid, ca_county, orange_single_day, Irvine_single_day, etc. to use in various data analysis.

To create a single day, I have used the last date function to get the latest trends in coronavirus cases.

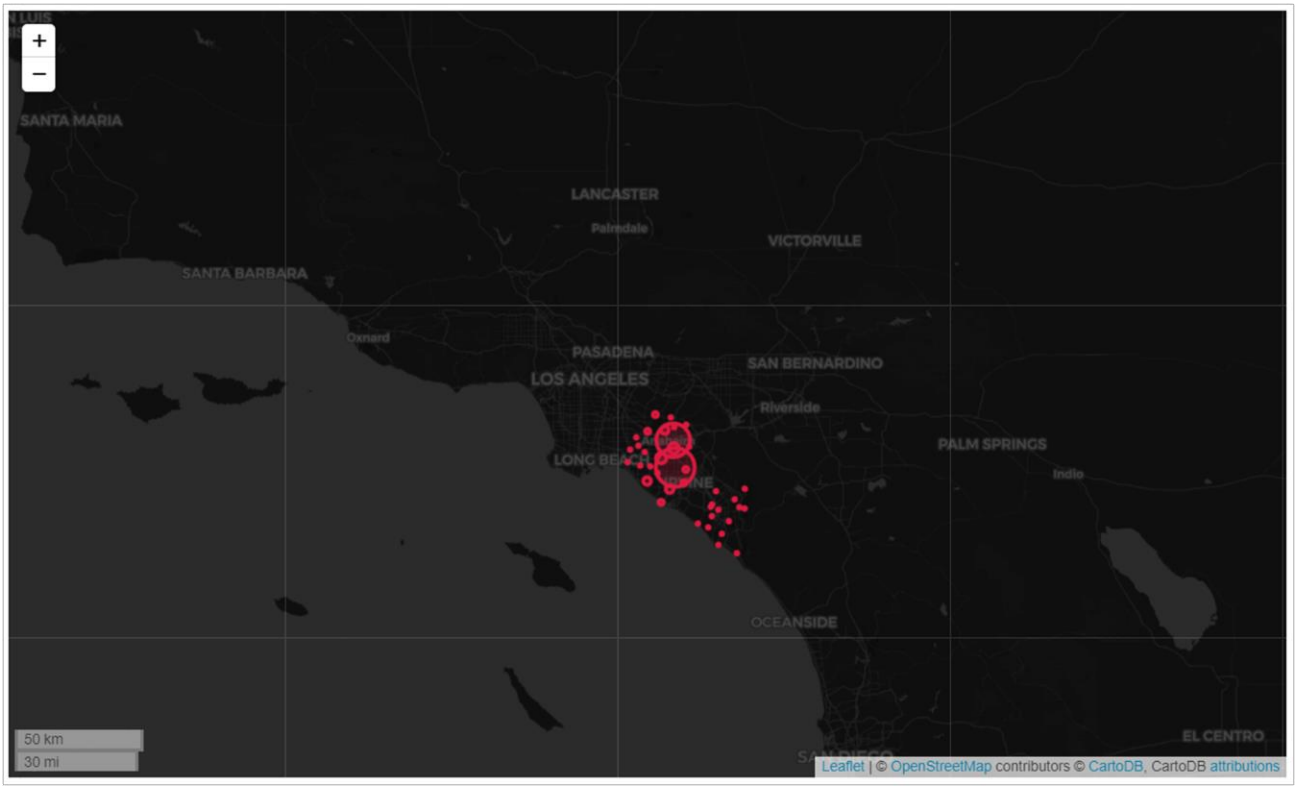
I am using Altair and folium to create an interactive map for California coronavirus data. You can zoom in, and each circle can provide information with a tooltip. We then use the python **folium** library to visualize the geographic details of Orange County and its places. I created a map of Orange with a superimposed circle on top using the latitude and longitude values represented with 'x' and 'y' to get the visual as below:

Center the map on a location of Irvine so that most data points are visible within the viewing area.



The CreateCircle () function create a circle of each place in the county with a number of confirmed cases with tooltip label. We turn off the label to make the map clutter-free. The circle dimension, with the help of Vega-Lite, determines the amplitude of confirmed cases.

Create a map with the center point and proper zoom level.



We create a new function as per our interest that will repeat the process above for all the neighborhoods in Orange County. This function will give us data of county with cases with the least and most cases. Here is a head () and tail () value of this data frame.

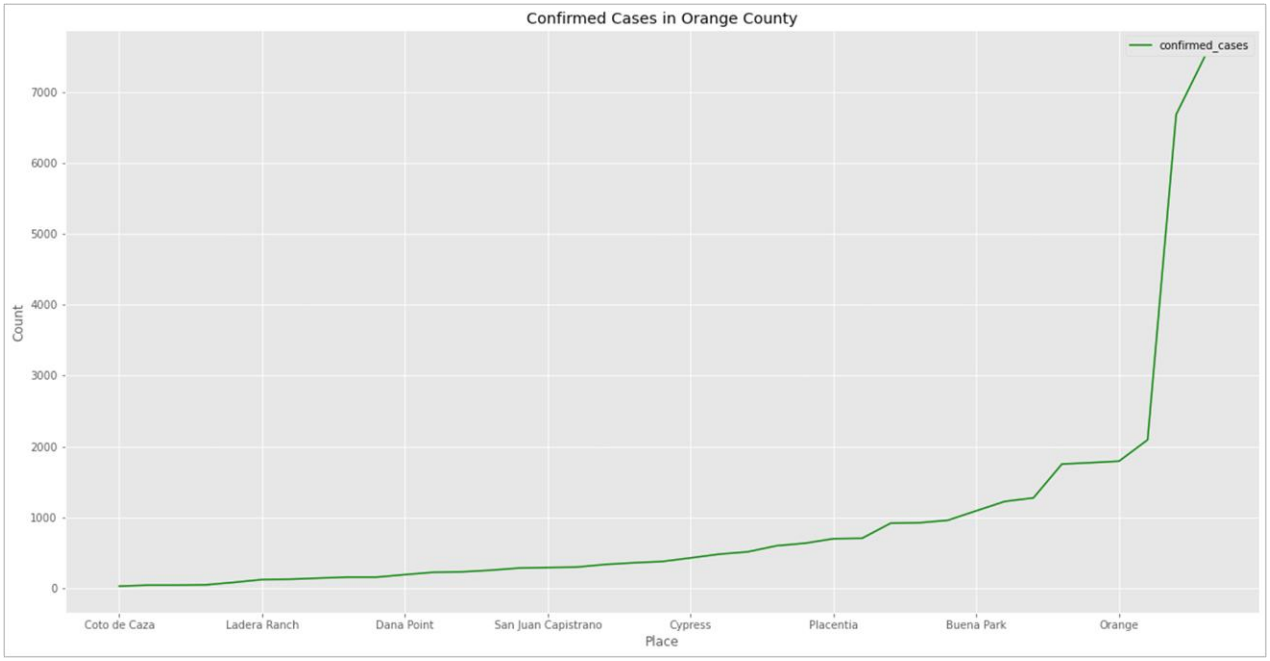
Orange county least affected area with coronavirus

	date	county	fips	place	confirmed_cases	note	x	y
455	2020-08-08	Orange	59.0	Coto de Caza	28	NaN	-117.587778	33.595833
479	2020-08-08	Orange	59.0	Rossmoor	29	NaN	NaN	NaN
484	2020-08-08	Orange	59.0	Silverado	40	NaN	NaN	NaN
488	2020-08-08	Orange	59.0	Villa Park	42	NaN	-117.822072	33.814006
469	2020-08-08	Orange	59.0	Laguna Woods	42	NaN	-117.725116	33.610170

The next data frame shows Orange county's most affected area with coronavirus.

	date	county	fips	place	confirmed_cases	note	x	y
482	2020-08-08	Orange	59.0	Santa Ana	7496	NaN	-117.881389	33.740833
451	2020-08-08	Orange	59.0	Anaheim	6682	NaN	-117.889722	33.836111
460	2020-08-08	Orange	59.0	Garden Grove	2094	NaN	-117.940639	33.773220
475	2020-08-08	Orange	59.0	Orange	1793	NaN	-117.887465	33.807614
459	2020-08-08	Orange	59.0	Fullerton	1771	NaN	-117.925205	33.871972

Orange County confirmed cases in different cities.



Analyze each neighborhood

We use Vega-Lite with folium to represent the data and find out the top five least affected cities and the top five most affected cities in the neighborhood.

Discussion:

The intent with which analysis was carried out was to find out areas with least and most affecting with coronavirus within California.

As we analyze the results section, we can analyze the different clusters like counties and places with each day of case count and see similar neighborhoods in different parts of the county. For example, if we compare the city of Santa Ana with another city Laguna Woods and see the difference in magnitude overlay on the map.

As seen in the table above, a person can determine which area is least infected, and if a person would like to book a hotel or to visit a plan, they can decide based on this analysis. This is just one example of how our data analysis can help people can trace the daily count and check the trend. The time series analysis can give an overview of a map of how coronavirus spread through in the area in the last six months or future.

Conclusion:

In such a unique time in recent history, there are many real-life problems or scenarios where data can be used to find out how we are dealing with the pandemic situation. As seen in the example above, data was used to cluster neighborhoods in Orange county or can be modified based on user choice. Similarly, data can also be used to make a calculated decision or make a trend for your city or county, which was not tracked through media.