Covid-safe areas in the US

Leonard Hu

27/06/2021

Introduction

As the covid vaccination rollout moves forward across the world, businesses are looking to expand and capture the resurgent market. Offices and stores are reopening at an increasing rate, and many expect much opportunity in the coming months. However, variants of covid, such as the highly infectious delta variant, are beginning to appear. Vaccinations do not completely immunise those who take it from covid's effects. And some members of the population are at-risk and/or medically unable to take the vaccine. Therefore, it is still wise to consider the risk of covid when opening a physical business location. This study will attempt to analyse coronavirus data for the entire United States to find the safest location in which to open a physical business location.

Lockdowns enforced during the pandemic have limited the ability of people to partake in physical activity. Fitness levels and general unhealthiness are major issues that are and will continue to be problems, even after herd immunity is achieved from the vaccine rollout. A location with easy access to local parks and exercise opportunities would serve to greatly distinguish a business and attract talent who would appreciate an emphasis on personal physical wellbeing.

Data - Sources

The data used will be the coronavirus data in the United States. The historical data used includes the cumulative number of cases and deaths for each location, each day. Three data sets are used:

- Daily data for the entire US
- Daily data for each state
- Daily data for each county

Source: https://github.com/nytimes/covid-19-data

Historical Data for US - https://raw.githubusercontent.com/nytimes/covid-19-data/master/us.csv

Historical Data for States - https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv

Historical Data by County - https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

In addition, geospatial data for the neighbourhoods of Los Angeles, CA was taken from the UCLA geoportal.

https://apps.gis.ucla.edu/geodata/sr_Latn/dataset/los-angeles-county-neighborhoods/resource/6cde4e9e-307c-477d-9089-cae9484c8bc1

Lastly, Foursquare data was used to find the types of venues located around each neighbourhood.

Data – Cleaning

The data for the cumulative number of Covid cases and deaths across the US, its states and counties needed to have its state names changed from its full names to their state codes, in order to work with Plotly. A dictionary linking state names with their state codes, courtesy of Roger Allen on Github, was used in this process.

Another issue with the data was that it was provided in cumulative form, making it less useful when attempting to visualise the amount of new cases in a certain time period. Furthermore, the most recent US census data for the population of its states was in 2019. In order to find the number of new cases of Covid as a percentage of a state's population, the population numbers needed to be estimated. This was done by calculating the average yearly growth of population for the years 2010 to 2019 and extrapolating the data to June 2021 using the average yearly growth.

The data provided by the UCLA geoportal included *all* neighbourhoods in Los Angeles, including some outlying neighbourhoods that would not be in consideration for a business. Therefore, neighbourhoods from certain regions such as Antelope Valley and the Santa Monica Mountains were removed to focus the results.

Finally, the geolocator used to attach latitudes and longitudes to neighbourhoods failed to find certain neighbourhoods, instead bringing up locations from outside Los Angeles, and even the United States. These erroneous locations were manually removed before the mapping and clustering process.

Methodology

In order to find the ideal location, the first point of interest is the state. Covid-19 has impacted the entire United States, but the severity of this impact and the response organised by local authorities are important factors to decide an ideal location. While the action taken by local authorities is a qualitative variable, we instead focus on quantitative variables such as the rate of daily cases, both absolute and as a percentage of each state's population.

The state names in the United State coronavirus data were recoded into state codes for compatibility with Plotly. Initial analysis shows that California, Texas, Florida and New York were some of the hardest hit states, with the highest number of total cases since records began. However, the data included everything since Jan 2020, the start of the pandemic, and it's reasonable to assume that states with higher populations would have a larger number of cases.

The data was restricted to the number of cases since April 2021, resulting in a 2 month period from 1 April to 1 June 2021. This revealed Florida, Michigan, New York, Pennsylvania and Texas as the states with the highest number of new cases in the 2 month period.

To obtain a clearer picture, new cases would need to be compared against a state's population. Updated population numbers are not available online, only data up to 2019. The population data for 2021 was this extrapolated using the average yearly rate of population growth from 2010 to 2019. The number of new Covid cases from 1 April 2021 to 1 June 2021, as a proportion of the state's total extrapolated population, was plotted on a map. Michigan was the state with the highest proportion of cases to inhabitants by a factor of 1.7, with Minnesota, Delaware, Colorado and Pennsylvania rounding out the top 5.

	state	% new cases
22	MI	2.397393
23	MN	1.409687
7	DE	1.397857
5	CO	1.383533
38	PA	1.357156

However, the focus of this assignment is to find the safest location. When sorted ascending, California is the state with the lowest number of cases as a proportion of its population, by a significant amount.

	state	% new cases
4	CA	0.302988
36	OK	0.364726
3	AR	0.366039
11	HI	0.367477
16	KS	0.397597

Thus, our state of choice is California. The next objective is to find the neighbourhood which suits the business and personal needs of employees.

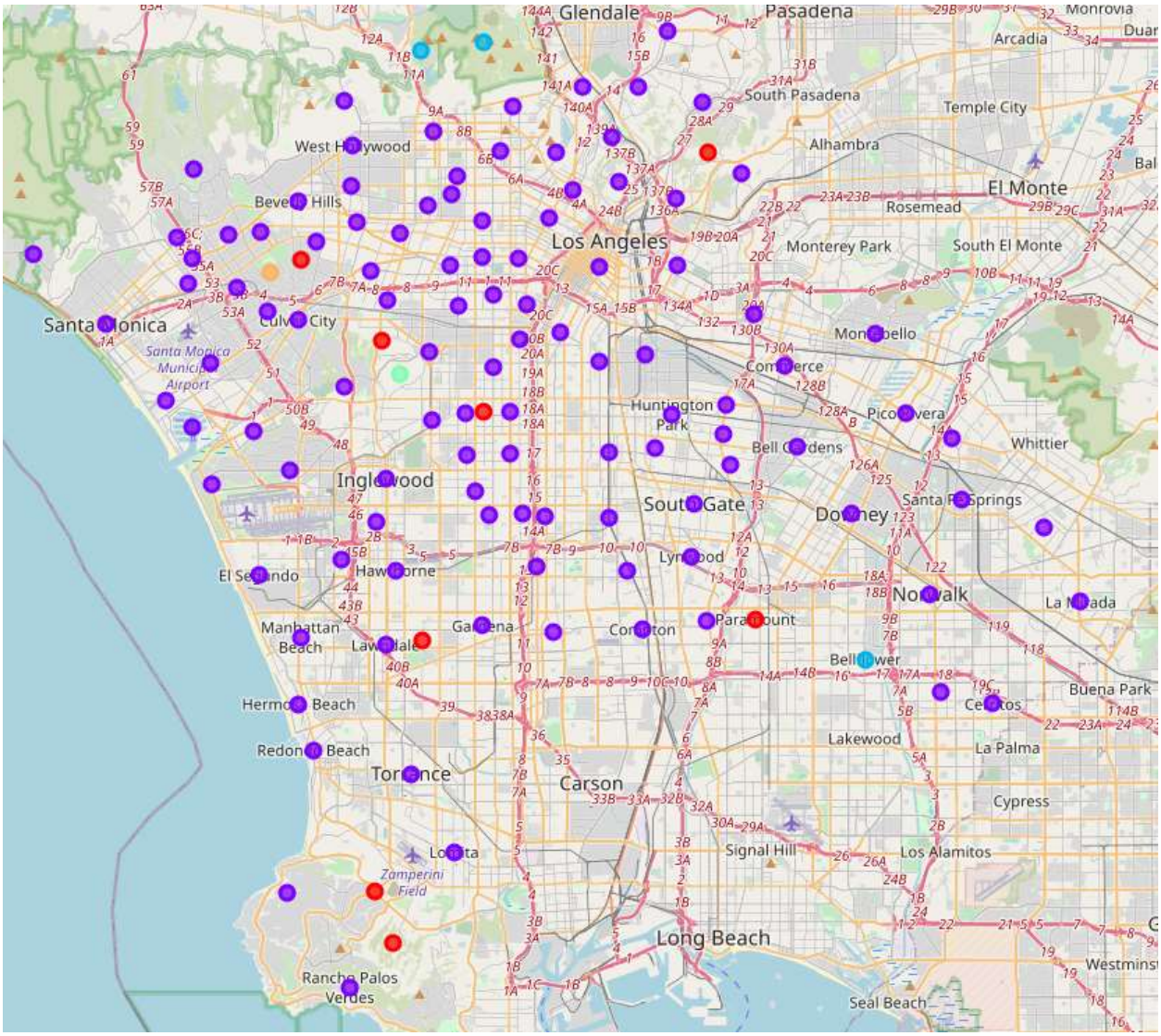
As Los Angeles is the largest city in the state, likely with the best amenities, the office will be opened there. Geospatial data for Los Angeles was taken from the UCLA geoportal, and latitudes and longitudes assigned to each neighbourhood using geolocator.

The venues for each neighbourhood were taken using the Foursquare API. These were then one-hot encoded to their respective neighbourhoods, and the mean of the occurrence values of each neighbourhood's venues were calculated into a dataframe. Then, these neighbourhoods were ranked according to their most common venue, and k-clustering applied to this data to determine and find clusters of neighbourhoods that share similar characteristics and common venues.

As a result of clustering, one cluster was identified as having ideal characteristics to solve the business' needs. The cluster 1.0 had Yoga Studios and Parks as their 1st and 2nd most common venues, ideal venues to have when attempting to build a business for employees to get exercise.

Results

Analysing the results of the k-clustering, cluster 1 seems to fit the needs of the business the most. With Parks and Yoga Studios among the most common venues in those neighbourhoods, as well as having Playgrounds and Farmers Markets, cluster 1 would be the perfect choice for businesses that want physical activity and exercise options close by.



[Cluster 1 marked in red]

The Los Angeles neighbourhoods that are in cluster 1 are:

Alondra Park, CA	Montecito Heights, CA
Baldwin Hills/Crenshaw, CA	Paramount, CA
Beverlywood, CA	Rolling Hills Estates, CA
Harvard Park, CA	Rolling Hills, CA

Discussion and Conclusion

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
1	Alondra Park, CA	Home Service	Park	Yoga Studio	Farmers Market
6	Baldwin Hills/Crenshaw, CA	Playground	Park	Yoga Studio	Farmers Market
14	Beverlywood, CA	Business Service	Park	Yoga Studio	Farm
47	Harvard Park, CA	Park	Yoga Studio	Farmers Market	English Restaurant
78	Montecito Heights, CA	Food	Park	Yoga Studio	Farmers Market
83	Paramount, CA	Mexican Restaurant	Park	Business Service	Burger Joint
92	Rolling Hills Estates, CA	Bank	Business Service	Farm	Park
93	Rolling Hills, CA	Business Service	Yoga Studio	Electronics Store	Escape Room

These neighbourhoods in cluster 1 have been identified as the most suitable for the business. The exact choice of location would depend on other factors as well, such as rent costs, proximity to competitors (in the case of businesses that directly compete for face-to-face customers), and public transport access. However, the use of machine learning has identified these neighbourhoods as having proximity to venues that would be desirable for a business that would like physical activity locations for their staff.

The modelling was not flawless – the k-cluster algorithm ended up clustering a large majority of data points into one cluster (cluster 2), while distributing a much smaller amount of data points across the other clusters. While this may be due to most neighbourhoods having very similar common venues, this is highly unlikely and further refinement of the k-clustering algorithm, or the use of another machine learning algorithm may return more accurate results.