Finding Coronavirus Hotspots

Shikhar Gupta

Introduction

It's no secret that we are currently living in a time of crisis, where the COVID-19 virus has become a global pandemic, forcing billions of people into quarantine for the last 3 months. Even with social distancing and isolation measures, the rates of infection continue to rise, and more and more people are getting the disease. Texas is one of the states in the U.S. that has implemented "reopening" measures, in which many nonessential businesses and recreational centers are allowed to open again. However, this process has led to a spike in infection rates as more people have started to go out again.

Thus, this project seeks to answer these essential questions:

1. *What counties in Texas have the highest infection rates?*
2. *What are the most popular types of venues for each county in Texas?*
3. *Is there a relationship between specific types of venues and infection rates?*
4. *Can we create a ranked list of venues that correlate to the highest infection rates based on this data?*

The audience for this project would be the general public, who would want to reduce the risk of coronavirus infection by avoiding particular types of popular venues when going out. Though there are many other factors that also come in to play in deciding infection rates for a county (such as testing rates, income, population, etc), this could give a good indication of places to avoid for concerned citizens who want to determine best practices for staying safe during the pandemic.

Data

Sources:

Foursquare API (from developer.foursquare.com) - The Foursquare API provides detailed and comprehensive location data that can be used to determine popular venues and details about those venues in a given area. This data was used for obtaining information about popular venues for each county in Texas and compared with the coronavirus case data in order to determine how coronavirus rates relate to different types of venues. The data was extracted through calls to the API within the Jupyter Notebook and received as a JSON file, which was used to convert into pandas data frame format for further processing.

Texas County Coronavirus Cases Data *(from dshs.texas.gov/coronavirus/additionaldata/)* - This comprehensive dataset is updated daily, and it gives the total population as well as the total number of confirmed cases for each county in Texas. This was used to determine the infection rate for each county, and it was processed to create choropleth maps for each county as well as correlating different types of venues to infection rates. This dataset is read in as an excel file, and this was converted to pandas data frame format for further processing.

Texas County Centroid Map

(from https://data.texas.gov/dataset/Texas-Counties-Centroid-Map/ups3-9e8m) - This dataset provides the central latitude and longitude data for each county, helpful for inputting data for search by the Foursquare API. This is given in CSV format and was converted into pandas data frame format for further processing.

Texas County Boundaries Map

(from http://gis-txdot.opendata.arcgis.com/datasets/texas-county-boundaries) - Finally, this data set is a GeoJson file that helped with visualizing the data through the Folium library, as well as creating choropleth maps when combined with coronavirus data.

Cleaning and Processing:

The data on coronavirus cases originally included many extra rows full of bibliography and credits, and the dataset also consisted of daily case counts starting from March, so many of the extraneous features needed to be removed. Furthermore, there was a total count of cases for all of Texas at the bottom that was also removed in processing. Thus, the only relevant features kept for the project were the county names, the total population for each county, and the current case totals.

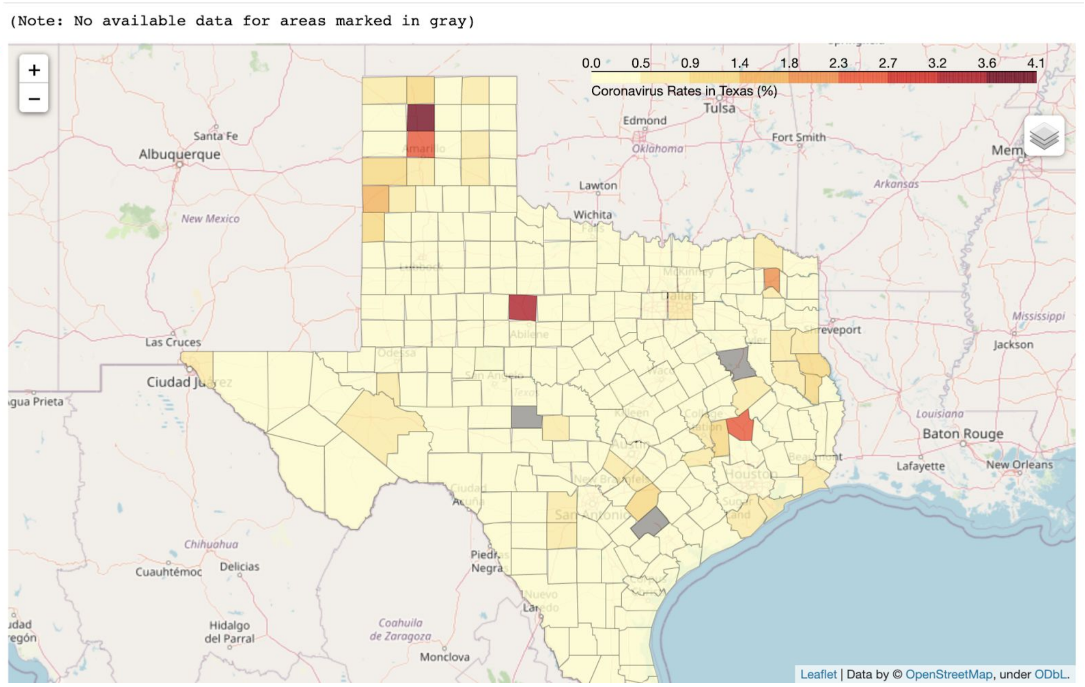
For the Texas County centroid map, the values for longitude and latitude, as well as county names, were selected in order to be fed into Foursquare API calls, however, in the original dataset, these were mislabeled. The column originally labeled as latitude was actually the longitude value, and the column labeled longitude was actually the latitude value. I renamed these columns in order to correctly represent the values.

Furthermore, the Texas Centroid dataset, the Texas Coronavirus Cases dataset, and the Texas County Boundaries map included most of the counties, but not all of them. In the beginning of the project notebook, there were calculations for about 254 counties, but due to some missing data between the three sources, the final number of counties ended up being 250 instead. This is more than sufficient to represent almost all of Texas, but it is important to note these omissions from the project.

Methodology

Calculating target variable

The original dataset for coronavirus cases did not include a column for infection rate for the county, so this was calculated by dividing the total number of cases by the population, and then this value was added alongside each county's statistics. The infection rate was then plotted on a choropleth map of Texas counties, generated using the Folium library and the GeoJSON files of Texas County Boundaries. This number was also multiplied by 100 for easier readability and for plotting purposes, since the percentages are fairly low.



```
(Note: No available data for areas marked in gray)
```

0.0 0.5 0.9 1.4 1.8 2.3 2.7 3.2 3.6 4.1
Coronavirus Rates in Texas (%)
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig 1. Coronavirus infection rate choropleth map of Texas counties

Finding popular venues for each county

Next, after cleaning the dataset for the Texas Centroids data set, I was able to use Foursquare API calls to retrieve popular venues for each county, and I searched in a radius of about 25km, enough to cover most of the area of a typical Texas county. In the search, I received back 9,517 total venues which were then put into a dataset and grouped by county, then converted to numerical values via one-hot encoding and frequencies for each were found. Then, this frequency was used to rank the top 10 most common/popular venues for each county.

Clustering analysis

Having found the ten most popular venues for each county, this data was then combined with the infection rates for each county and put into a data set. The venue data was again processed with one-hot encoding techniques, and then the entire dataset was normalized using the min max scaler method. K-means was then chosen as the method for clustering, and the ideal k value was found by running the K-means algorithm on many different values for k and minimizing the distortion using Euclidean distance between the cluster points and the centroids. Using the elbow method, I concluded that the ideal K value was 8, and this value was used to create 8 clusters, which were then analyzed and superimposed onto a cluster choropleth map of Texas counties.

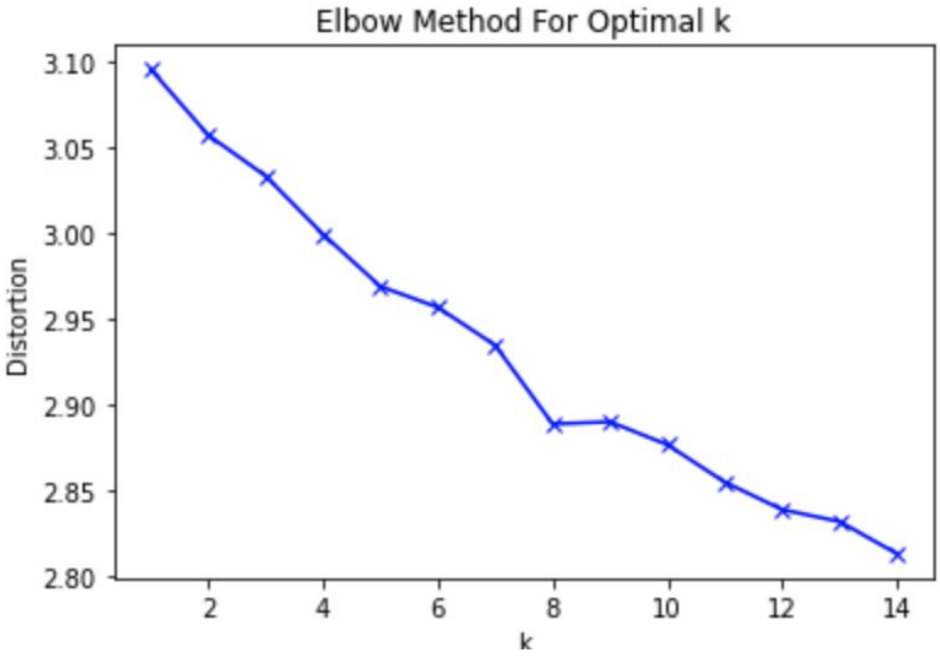


Figure 2. Distortion graph for K-means clustering

(Note: No available data for areas marked in gray)

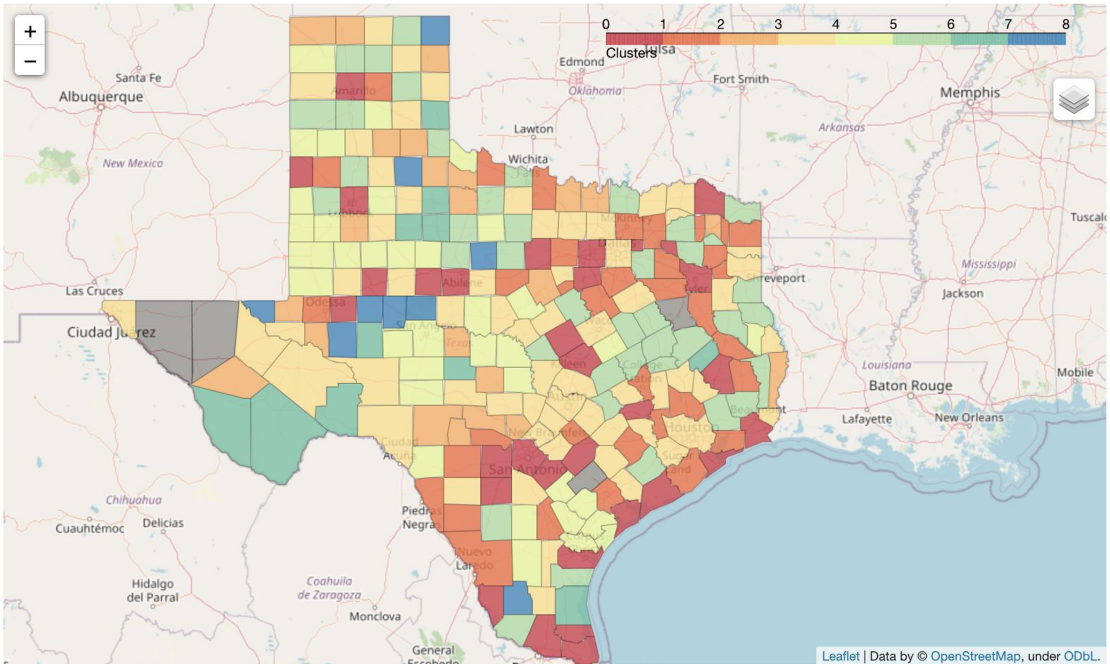


Figure 3. Cluster choropleth map of Texas counties

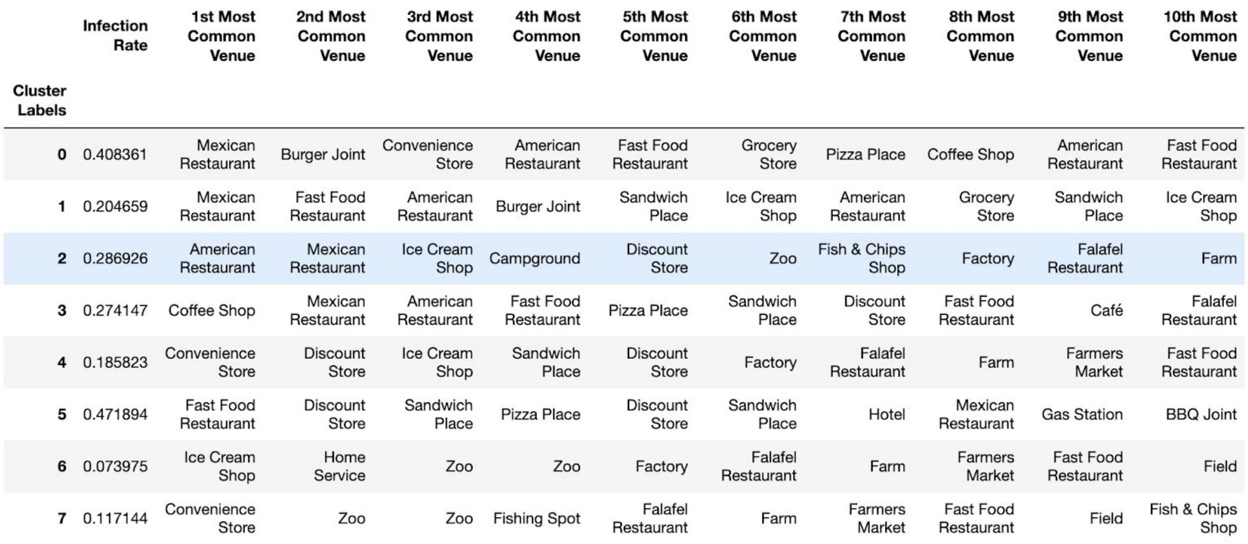
Ranking venues

Finally, a comprehensive ranking of different types of venues was done based on their computed contribution to coronavirus infection rates. This metric was calculated via a weighted system where the ranking of the venue for each county (1st most popular to 10th most popular) was weighted on a scaled of 0.1 to 1 (1st having a weight of 1), and then this weight was multiplied by the normalized infection rate for that county (from 0 to 1). This ensured that the most popular venues in high infection rate counties counted the most towards the point total while less popular venues in lower infection rate counties didn't count as much. The total score was averaged by dividing by the total frequency for each venue type. The resulting dataset was also purged of venues that didn't appear at least 5 times, and the venue, "intersection", was removed from the dataset since it logically would not be feasible as a candidate. Then, using the final scores, a comprehensive ranking was created of types of venues.

Results

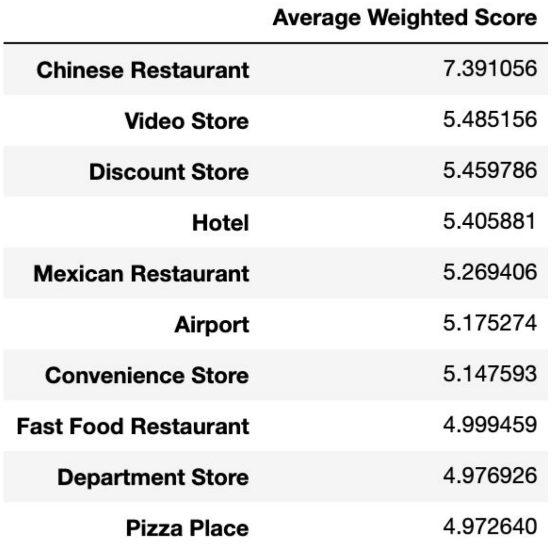
Clustering

The results for the 8 clusters are shown below. For each cluster, the infection rate and the most common venue for each category (1st most common, 2nd most common, etc) are shown. Through observation of the clusters, I noted that the clusters with a higher overall infection rate (such as cluster 0 and 5) also tended to have more stores and restaurants as popular venues while clusters with lower rates, like clusters 6 and 7, tended to have less restaurants and stores, more often having venues like farms and zoos.



Cluster Labels	Infection Rate	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0.408361	Mexican Restaurant	Burger Joint	Convenience Store	American Restaurant	Fast Food Restaurant	Grocery Store	Pizza Place	Coffee Shop	American Restaurant	Fast Food Restaurant
1	0.204659	Mexican Restaurant	Fast Food Restaurant	American Restaurant	Burger Joint	Sandwich Place	Ice Cream Shop	American Restaurant	Grocery Store	Sandwich Place	Ice Cream Shop
2	0.286926	American Restaurant	Mexican Restaurant	Ice Cream Shop	Campground	Discount Store	Zoo	Fish & Chips Shop	Factory	Falafel Restaurant	Farm
3	0.274147	Coffee Shop	Mexican Restaurant	American Restaurant	Fast Food Restaurant	Pizza Place	Sandwich Place	Discount Store	Fast Food Restaurant	Café	Falafel Restaurant
4	0.185823	Convenience Store	Discount Store	Ice Cream Shop	Sandwich Place	Discount Store	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
5	0.471894	Fast Food Restaurant	Discount Store	Sandwich Place	Pizza Place	Discount Store	Sandwich Place	Hotel	Mexican Restaurant	Gas Station	BBQ Joint
6	0.073975	Ice Cream Shop	Home Service	Zoo	Zoo	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field
7	0.117144	Convenience Store	Zoo	Zoo	Fishing Spot	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Fish & Chips Shop

Fig 4. Results of cluster analysis



	Average Weighted Score
Chinese Restaurant	7.391056
Video Store	5.485156
Discount Store	5.459786
Hotel	5.405881
Mexican Restaurant	5.269406
Airport	5.175274
Convenience Store	5.147593
Fast Food Restaurant	4.999459
Department Store	4.976926
Pizza Place	4.972640

Fig 5. Final venue rankings

Ranking

The final rankings and weighted average scores are shown above. The higher the score, the more that type of venue is predicted to contribute to the coronavirus infection rate, thus making it more likely to be a "hotspot". Thus, it was found that Chinese Restaurant was considered to be the highest risk venue, and other stores and restaurants were also included in the hotspot list, confirming the observations made during the clustering stage.

Discussion

Implications

The results and the data obtained from this project suggest there do exist relationships between the types of venues present in certain Texas counties and infection rates. The clusters with higher infection rates also seemed to be more urbanized, suggested by the most frequent venues including restaurants, convenience stores, and fast food restaurants. Conversely, the clusters with lower infection rates seemed to be more rural in nature, including venues such as farms. On the choropleth map of clusters, many of the clusters are grouped together spatially, suggesting a similar rural or urban landscape as well. This would merit further study into how urbanization affects the spread of disease.

External Factors and Validity of Results

The results drawn from this experiment are purely exploratory in nature and cannot comprehensively be used to draw a statistical correlation between types of venues and the coronavirus infection rate. This is because there are many external factors that would also affect infection rates such as testing numbers, numbers of asymptomatic carriers, income level, etc. that would most likely also play a role in determining infection rates. The ranking of venues is suggestive as well and does not necessarily mean that those types of venues are inherently more dangerous. However, this study does point to a possible relationship between the frequency of certain types of venues and infection rates, as can be seen from the cluster analysis and ranking results. However, further statistical testing and removal of external factors would be needed to draw cohesive conclusions.

Conclusion

Thus, this study gives a good insight into one of the many under researched factors that may play a role in determining infection rates in a region. This study is even more telling when considering the fact that most of these venues have reopened due to Texas' reopening policy. Further research in this topic could include a national study, and also could consider the increase in the number of cases over time rather than just the total number of cases. This study could be used as a factor in determining preventative measures for slowing the spread of the disease, but further statistical testing and correlation analysis can also be done and compared with other external factors to measure the significance. A worthwhile study to do in the future would also be to compare the results from this project with a similar study done in an area where reopening has not been implemented to see the differences in infection rates and venues.