Capstone Project – The Battle of Neighborhoods

by

www.coursera.org

"COVID-19: Numbers & Rates in TX State"

by

Ing. Luis Javier García Murillo

Applied Data Science Capstone

by

IBM | Alex Aklson

May 12th, 2020

Monterrey, NL, Mexico

INDEX

Keywords:

COVID-19, China, United States, Texas, Infected rate, Death rate, Recovery rate, EDA, Foursquare, County, Population, Cinema.

I. Introduction

During this year, the world has witnessed a new virus called Coronavirus-19 (COVID-19). The outbreak began in late December last year in a city from China called Wuhan. This virus spreads in three possible ways; direct contact, droplet spray in short range or aerosol in long-range. Thanks to its high propagation capacity, during the first quarter of this year, and the high level of economic activity in China. The virus was able to spread in other countries such as Europe, Italy and the United States, respectively.

In this report we will talk about how this virus has infected the North American population, specifically in the state of Texas, as the state closest to the border with Mexico and therefore the main source of infection for our country. The ten most infected counties in the state of Texas will be revealed, important data will be shown such as; the infected rate, the death rate, and the recovery rate by county.

In addition, an Exploratory Data Analysis (EDA) will be conducted with information from Foursquare to seek to find a relationship or cause for these 10 counties to become the most infected states in Texas. To conclude, we will discuss the results of the analyzes and seek to reach a coherent conclusion for all readers.

II. Data sources / Pre-processing

For the preparation of this report, we set out to use the skills that this course taught us. We started by looking for the COVID-19 cases as up to date as possible and that it was separated by county in the State of Texas, USA. We found a table inside the Wikipedia web page [1] and we proceeded to extract and clean it using the Pandas library as we were taught in this course. The result of the data frame can be seen in Figure 1.

Out[2]:

County	Cases	Deaths	Recoveries
Anderson	23	0	0
Andrews	19	0	0
Angelina	46	0	0
Aransas	2	0	0
Archer	0	0	0
Armstrong	2	0	0
Atascosa	16	1	0
Austin	12	0	0
Bailey	0	0	0
Bandera	6	0	0

Figure 1. COVID-19 numbers per county in Texas. (source: Wikipedia [1])

Subsequently, we had to find the total population by county to obtain the indices mentioned in the introduction to this report. A table was found within the official Texas Demographics page [2] and we proceeded to extract it, clean it and order it to be able to join it with our first table (fig. 1). The extraction result can be seen in Figure 2 below.

Out[3]:

	County	Population
56	Anderson	57863.0
130	Andrews	17818.0
43	Angelina	87607.0
102	Aransas	24763.0
171	Archer	8789.0
235	Armstrong	1916.0
67	Atascosa	48828.0
94	Austin	29565.0
189	Bailey	7092.0
111	Bandera	21763.0

Figure 2. Population per county in Texas. (source: Texas Demographics [2])

Then we proceeded to search the latitude and longitude data for each of the counties in Texas. The research was not easy, but we found in Gaslamp Media [3] our information. First, a list was downloaded and then loaded on the Jupyter platform to be able to upload it to our notebook. The result is shown in Figure 3. The latter to be able to relate our data to some Foursquare data that we seek.

Out[4]:

	county	latitude	longitude
0	Anderson	31.794191	-95.661964
1	Andrews	32.370377	-102.523255
2	Angelina	31.285984	-94.577084
3	Aransas	28.058922	-97.093640
4	Archer	33.615708	-98.687302
5	Armstrong	35.049184	-101.409336
6	Atascosa	28.926371	-98.524247
7	Austin	29.849283	-96.313271
8	Bailey	33.898805	-102.797253
9	Bandera	29.729603	-99.191141

Figure 3. Latitude & Longitude per county in Texas. (source: Gaslamp Media [3])

In order to begin carrying out our analyzes, we proceeded to clean up our final data frame and for this, all counties where there are no records of cases infected with the virus were removed. Then, they were ordered to obtain the first 10 counties with the highest number of COVID-19 cases registered. The result of our final data frame with some columns calculated from the data in Figs. 1 and 2 are shown below in Figure 4.

Out[18]:

	County	Cases	Deaths	Recoveries	Population	latitude	longitude	Infected (%)	Death Rate (%)	Recovery Rate (%)	Cinemas
0	Harris	8176	179	3016	4602523	29.833990	-95.434241	0.18%	2.19%	36.89%	9
1	Dallas	6123	145	2511	2586552	32.767268	-96.777626	0.24%	2.37%	41.01%	5
2	Tarrant	3745	104	780	2019977	32.771419	-97.291484	0.19%	2.78%	20.83%	7
3	Travis	2171	65	713	1203166	30.326374	-97.771258	0.18%	2.99%	32.84%	17
4	Bexar	1920	57	976	1925865	29.437532	-98.461582	0.10%	2.97%	50.83%	13
5	Fort Bend	1404	40	236	739342	29.525461	-95.771651	0.19%	2.85%	16.81%	2
6	El Paso	1348	33	685	837654	31.694842	-106.299987	0.16%	2.45%	50.82%	8
7	Potter	1179	17	158	120899	35.401475	-101.895089	0.98%	1.44%	13.40%	0
8	Denton	946	25	440	807047	33.195872	-97.116282	0.12%	2.64%	46.51%	6
9	Collin	939	29	587	944350	33.152417	-96.621427	0.10%	3.09%	62.51%	12

Figure 4. Top-10 Counties in Texas with the greatest number of COVID-19 cases registered.

III. Methodology

During the development of the project, many situations were presented in which the skills learned in past modules of this course were put to the test.

The objective of this report is to look for a correlation between the number of cases of COVID-19 and the number of cinemas that exist in each of the 10 counties with the highest number of registered infected, since these places are always full of people. To achieve this, three different sources were investigated to obtain the final data frame (Fig. 4).

In order to visualize the data in a better way, bar graphs were made. First, we needed to see how the numbers of cases looked like compared to deaths and recoveries. In Figure 5, we can observe that.

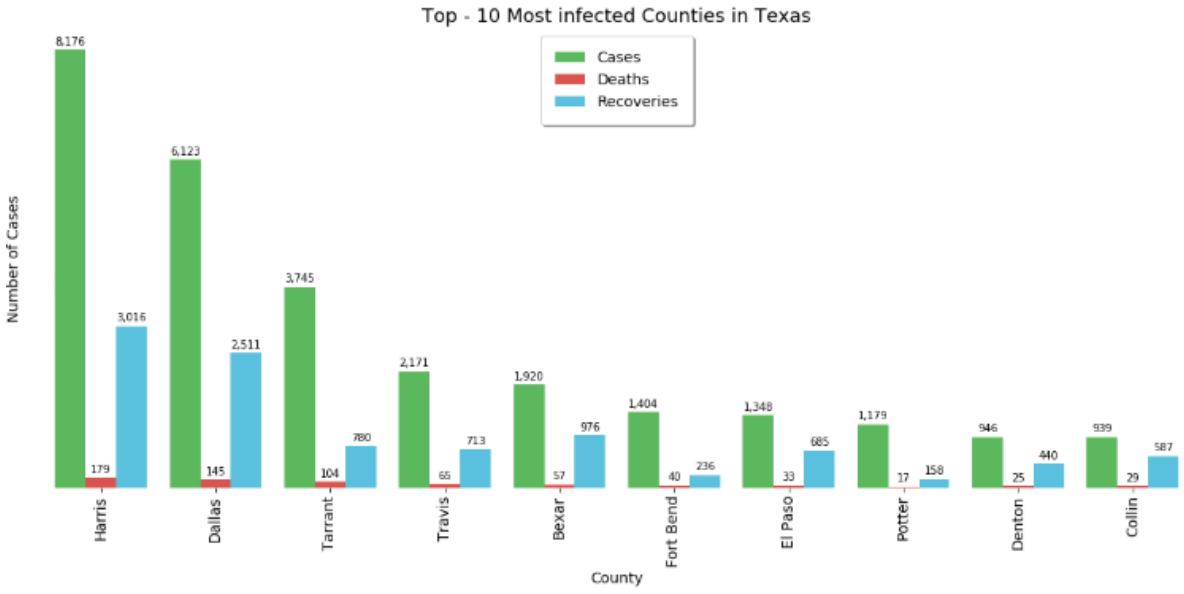

Figure 5. Top-10 Counties in Texas bar chart of the COVID-19 cases. (updated: 14/05/20)

As we can see in Fig. 5, the counties with the highest number of cases are: **Harris, Dallas & Tarrant** with almost **65%** of all registered cases in the state of Texas. A total of **27,951** COVID-19 cases is the total registered in just the Top-10 counties of Texas State. This represent the **68%** of the cases in the whole State. Let's see in Figure 6 how near these counties are from each other.

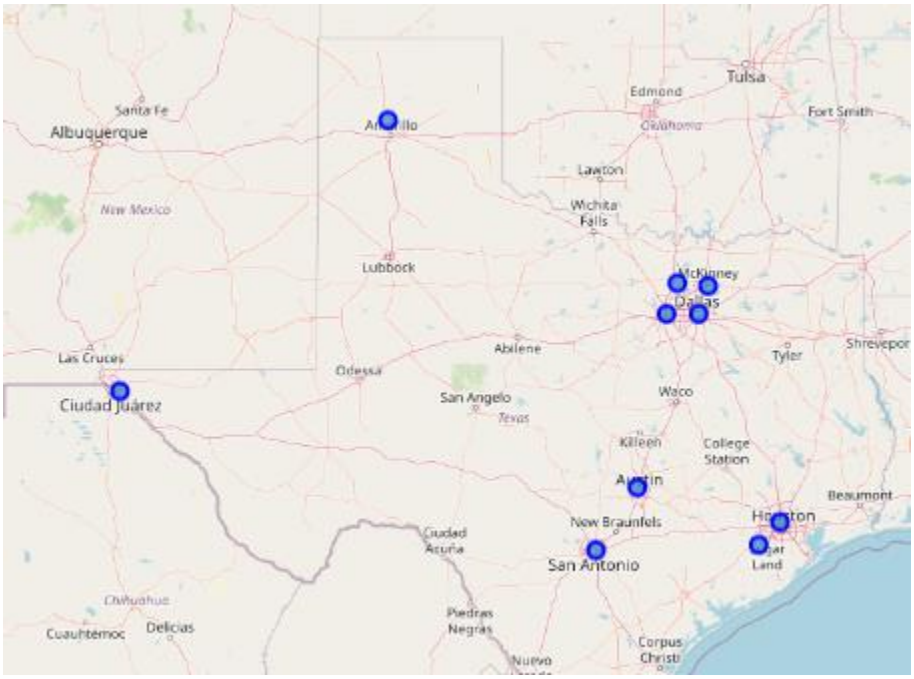

Figure 6. Top-10 Counties in Texas with a greater number of cases.

Now, let's see what these numbers look like compared to the number of inhabitants in each county. In Figure 7, we can see the percentage of infected by the virus by county.

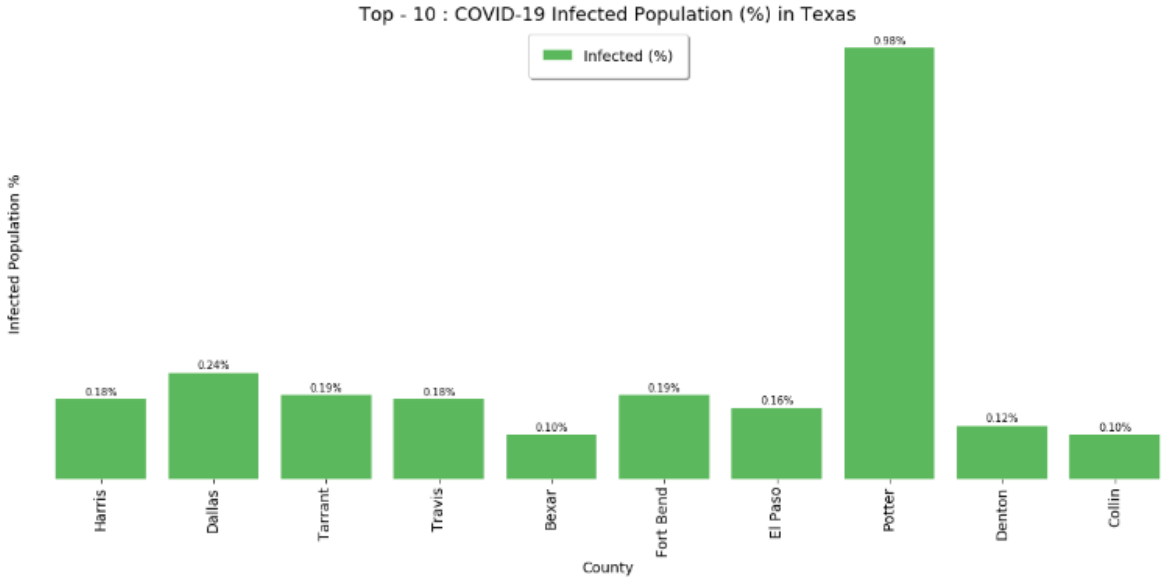

Figure 7. Infected Rate (%) of Top-10 Counties in Texas with a greater number of cases.

We can see that these numbers are not very significant compared to the entire population since in no county does it even reach **1%** of its population. But this does not mean that we can lower our guard. As we can see in Figure 8, the average death rate for the 10 counties is **2.57%**, which is equivalent to **72** deaths on average per county.

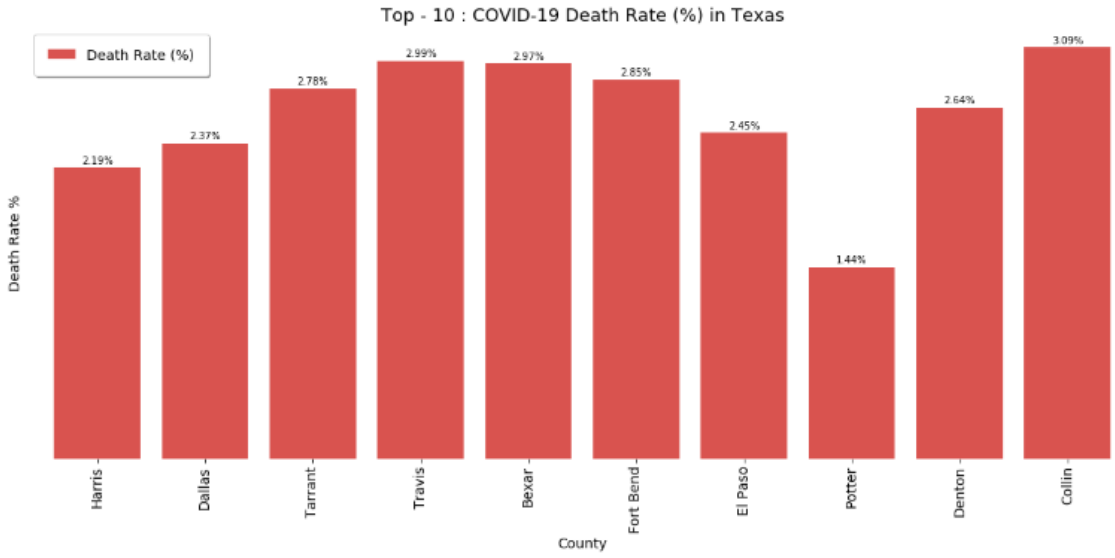

Figure 8. Death Rate (%) of Top-10 Counties in Texas with a greater number of cases.

But not everything is bad, if we look at the average recovery rate is **37.4%** equivalent to **1,041** people recovered by county, compared to the deaths it does not feel so bad anymore. Figure 9 shows this with the Top-10 counties.

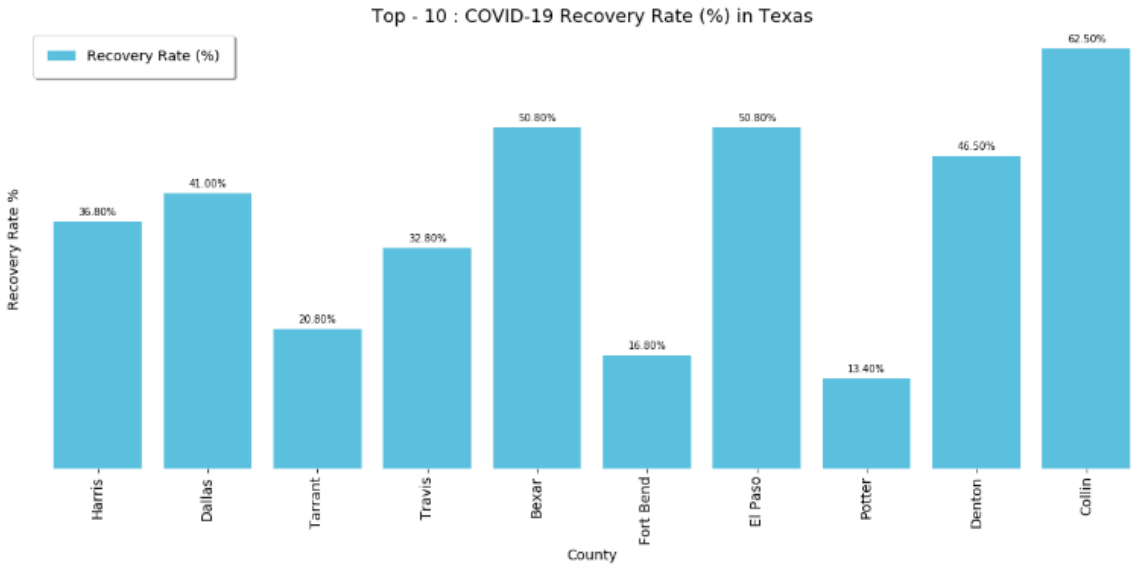

Figure 9. Recovery Rate (%) of Top-10 Counties in Texas with a greater number of cases.

In summary, the total number of cases in Texas state is **40,879 (0.14%)**. The number of registered deaths is **1,131 (0.004%)**. The number of cases recovered is **16,287 (0.05%)**. These numbers are based on Texas population of **27,769,133** people.

IV. Results & Discussion

To find out if our hypothesis is significant, we proceeded to perform a correlation analysis between the number of cinemas in the county (information extracted from Foursquare [4]). In Figure 10, we can see the information from Foursquare. In Figure 11 and 12 you can see the correlation analysis that was performed.

Out[17]:

	County	Cinemas
0	Harris	9
1	Dallas	5
2	Tarrant	7
3	Travis	17
4	Bexar	13
5	Fort Bend	2
6	El Paso	8
7	Potter	0
8	Denton	6
9	Collin	12

Figure 10. Number of cinemas per county in Texas (source: Foursquare [4])

As we can see in Figure 11, the correlation that exists between these two variables is not significant but shows a **weak positive correlation** due to the value obtained from (**0.017**, very close to 0).

Out[27]:

	Cases	Cinemas
Cases	1.000000	0.017463
Cinemas	0.017463	1.000000

Figure 11. Correlation table between Cases & Cinemas.

A better way to look at that relationship is with a scatter plot to look at the trend of the data and its pattern. In Figure 13, we can see that in fact the data does not follow a specific pattern and therefore the correlation is not entirely predictable, but it does exist.

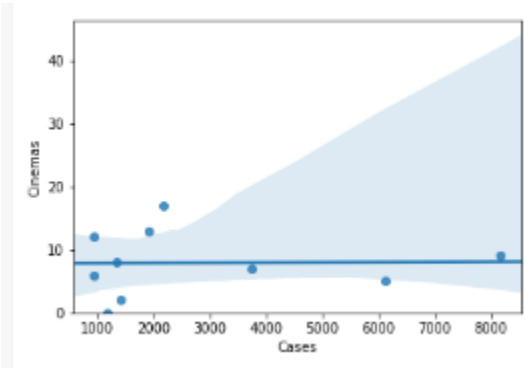

Figure 12. Scatter plot showing correlation between Cases & Cinemas.

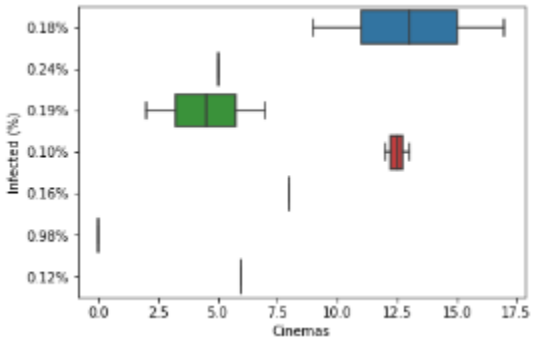

Figure 13. Box plot showing distribution between Infected Rate (%) & Cinemas.

A box plot was made but now using the infected rate to see how scattered the data is from each other. In Figure 13, it can be seen that there is no significant overlap between the data, and therefore the number of cinemas in a county may be one of the many causes where the virus could spread, although the correlation between the data is low.

V. Conclusion

To conclude, we can mention that our hypothesis was correct, but not as expected. The number of cinemas within a county can be one of many reasons why the virus spread. A correlation test was performed and the results were as follows; Since the p-value isn't <0.001, the correlation between Cinemas and Cases isn't statistically significant, and the linear relationship is quite weak (**0.0174**, close to 0), but this means that it is one of many other causes for which spread the virus, such as; the number of gyms, international arrivals per day, number of public parks, among others.

VI. Bibliography

[1] COVID-19 pandemic in Texas. (2020, May 15). Retrieved from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Texas#Statistics

[2] Texas Counties by Population. (n.d.). Retrieved from https://www.texas-demographics.com/counties_by_population

[3] Yin, Heidersays, T., Frenettesays, J., O'Shaughnessysays, M., Wisesays, D., Ochoasays, A., … Mohanrajsays, S. (2020, January 19). Download: Zip Code Latitude Longitude City State County CSV. Retrieved from https://docs.gaslamp.media/download-zip-code-latitude-longitude-city-state-county-csv/

[4] Foursquare search: “Cinemas” in Texas. Retrieved from https://developer.foursquare.com/