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

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[5]:

County	Cases	Deaths	Recoveries	Population	latitude	longitude	Infected (%)	Death Rate (%)	Recovery Rate (%)
Harris	6838	133	0	4602523	29.833990	-95.434241	0.15%	1.95%	0.00%
Dallas	4133	111	0	2586552	32.767268	-96.777626	0.16%	2.69%	0.00%
Tarrant	2584	75	0	2019977	32.771419	-97.291484	0.13%	2.90%	0.00%
Travis	1756	52	378	1203166	30.326374	-97.771258	0.15%	2.96%	21.53%
Bexar	1613	48	531	1925865	29.437532	-98.481582	0.08%	2.98%	32.92%
Fort Bend	1187	28	163	739342	29.525461	-95.771651	0.16%	2.36%	13.73%
El Paso	1029	22	486	837654	31.694842	-106.299987	0.12%	2.14%	47.23%
Potter	818	9	30	120899	35.401475	-101.895089	0.68%	1.10%	3.67%
Denton	806	22	392	807047	33.195872	-97.116282	0.10%	2.73%	48.64%
Collin	804	22	533	944350	33.152417	-96.621427	0.09%	2.74%	66.29%

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