Diljit Bal

Ead 502

9/4/2022

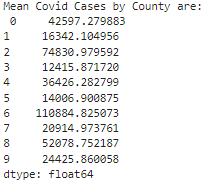
DS 5003 Exam 1

Question 1

The manner in which the data was collected was stated in the Texas Health and Human Services website (Specifically this link here <https://www.hhs.texas.gov/services/health/coronavirus-covid-19/texas-covid-19-case-count-vaccination-data>) On the site it explains that the following Covid-19 data is the information that was reported by facilities operated and/or regulated by HHSC. The data is consistently updated each weekday by 3 P.M. Central time and is provisional and subject to change. When looking at the Data Life Cycle and trying to pinpoint where Texas HHS is, since they are providing me already researched and defined data, they have already gone through the initial steps of Discovery and Data Preparation. The point of the cycle I got involved in was Model Planning and then Model Building. Model Planning took place when I first received the data and was determining how to interpret and analyze the data. Once I decided that the measure of centrality I wanted to focus on for this data set was the Mean, or in other words the average, I then moved on to the Model Building phase which consisted of me beginning to write the code script and then transitioning the data from being information on a table to a physical representation of said data. With all that said, the effects that the discovery of a new variant has are evident in that there will be a spike of cases across all counties typically during/around the time of said variant discovery, and since this is a on-going study (As listed in the HHS website), the measures of centrality are subject to consistent change.

Question 2

For this analysis, I generated the basic measures of centrality, so Mean, Median, and Mode, across all the counties as well as generating the Mean across all days. Since I was tasked with specifically monitoring counties that exceed a population size above 750k, I first modified my dataset/csv file only to contain those said counties, which ended up being the following: 0. Bexar, 1. Collin, 2. Dallas, 3. Denton, 4. El Paso, 5. Fort Bend, 6. Harris, 7. Hidalgo, 8. Tarrant, and 9. Travis. After running the updated CSV file, I ran the measures of centrality and received the following values:



Table

Description automatically generated with low confidence

Table

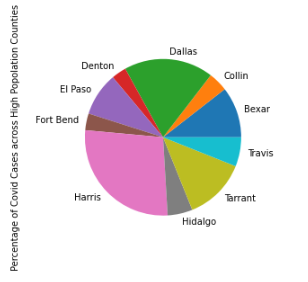
Description automatically generatedTable

Description automatically generated with medium confidence

Once I was at this point of analyzing the data, I then began to brainstorm which dataset I wanted to continue using on to the Model Building phase. I automatically realized that the mode, while a measure of centrality, didn’t really help with providing much information of overall covid cases occurring, so I marked that out. Median provided a great measure, the center point, and while it was a great and easy to see representation of which county was having more cases, I still thought that there would be a better method of viewing the average number of cases that occurred through the 11-month time period in which the dataset I received measured. This led me to the conclusion of using the Mean as the primary measure of centrality I wanted to observe through my analysis of Covid Cases across these multiple counties. Having the Mean of covid cases not only allowed to see the average number of cases that occurred within this time frame, but it also allowed us the ability to compare in terms of percentages which county with populations above 750k suffered worst from the outbreak of Covid-19 relative to the others. By having the Mean of covid cases that occurred daily, we can see the rise of Covid-19 throughout high-population counties in the 11-month time period.

Chart, bar chart

Description automatically generatedQuestion 3 – Visualization



Chart, line chart

Description automatically generated

Question 4

Since I was assigned the data subset of counties with a population above 750k, as stated in question 2, I trimmed the original csv. file down to only the counties that apply to my constraints. After that, I made a dataset the contained the mean values of cases in county and the cases on a certain day. After turning them into a physical representation/graph, it is easy to see how the counties rank on severity of cases as well as identify the solid & steady rise of Covid-19 throughout high population counties in Texas. As evident from the bar chart, it is seen that Harris County significantly outpaces the rest in covid case occurrences, followed by Dallas County, and then Tarrant. After viewing this information, I realized it had the potential to be separated into percentages based out of the total cases that occurred throughout theses 10 counties. This led to the formulation of the pie chart above and showcases the percentages of cases as a whole and how each county builds to that total percentage. Once you view the information here it is especially easy to quickly identify that Harris County accounts for nearly 25% of the total cases of Covid-19, followed by Dallas County which also nearly counts as 25% of total cases in high population counties in Texas. You can also see how Bexar and Tarrant counties are nearly the same in percentages (In the Bar Chart however it is more easily distinguishable that Bexar is less than Tarrant) and then after those two counties the percentage of cases per country tends to equalize to smaller percentages of the total number of cases. Since those two graphics/charts covered the information derived from calculating the Mean of Covid-19 cases that occurred during a 11-month timeframe, you can see that comparing this information to the data of county populations (TexasPopByCounty2021.csv), you can infer that Covid-19’s overall spread is affected by large population size, so in short, if the population is higher, the amount of Covid cases is bound to follow. Comparing the data to the information brings up an interesting caveat as well. You can see that while El Paso has the lowest population among the high population counties, it has a higher occurrence of Covid cases compared to some of the counties that have higher population. Perhaps their Covid-19 restrictions weren’t as strict/effective as other high population counties. Moving on, after calculating the Mean of Covid Cases across a county, I next wanted to interpret the information of how many cases were occurring daily and its growth throughout the 11-month period. This led me to creating the line chart above. Viewing the line chart, it’s easy to interpret the growth of cases within high population counties in Texas. You can see early quarantine was effective to limiting the spread of Covid-19, but once the summer months came (Post 5/29/2020), you see a sharp bump occur and after that the spread of Covid pretty much grew at a steady rate that was stronger than the amount it was growing during the first 2 months. With all this information in hand, we can start to put it into play a derive knowledge from it. I slightly stated some of the key points I derived already in explaining the information, but as you can see, in high population counties Covid is certainly growing in a steady pace. Besides the outlier of El Paso, we can infer that the higher the population of a county, the greater the size of those infected with Covid-19, thus increasing an individual’s chances of contracting the virus. This information also reveals that if high-population counties enforce Covid-19 restrictions and the populous does as well, there can be a reduction in the number of cases. I came to this conclusion based on the outlier of El Paso County. While on average the higher the population the more cases, there can still be less populous counties that still have a higher rate on Covid-19 transmission based on several variables, that while not listed here, can be determined by further evaluating the spread of Covid-19 in that specific county. All in all, the wisdom I receive from evaluating all of this is the following: No matter what the population size is, Covid-19 is growing in a steady rate across all counties in Texas. It would be best to follow procedures and other Covid-19 guidelines in order to insure that you will not contract the virus within any county, because even if it has a low populous there is still a chance you can receive it from an infected individual.