Analysis of Diseases of Significant Growth in the United States by State or Special Region

Patrick Cook

7/31/2021

Table of Contents

Project Overview

Project Topic

This project analyzes state level weekly cause of death data available from the Centers for Disease Control (CDC) to identify specific diseases with the highest growth in deaths so that it may be used to direct resources to reducing the number of deaths in the disease category.

Justification for Organizational Need

In 2019, the United States spent $3.8 trillion on healthcare accounting for 17.7% of Gross Domestic Product and the spending is projected to reach $6.2 trillion by 2028 (Centers for Medicare & Medicaid Services, 2020). Despite spending more on healthcare than any other modern country, the United States has worse outcomes (Tikkanen & Abrams, 2020). Worldwide, 20 to 40% of total healthcare spending is considered wasted spending (Sorato, Asl & Davari, 2020). In the United States, it is estimated that 25% of healthcare spending is wasted spending (Shrank, Rogstad & Parekh, 2019). Using the projected amount of healthcare spending for 2028, this would amount to more than $1.5 trillion dollars in wasted spending or the equivalent to more than $4,200 per person in the United Stated based on the U.S. Census Bureau's International Data Base Projections for 2028 (U.S. Census Bureau, n.d.). Using historical health data can reduce these wasted expenditures by predicting trends in diseases and ensure resources are available and targeted to what is needed at the time. This will further reduce expenditures by reducing wasted labor hours and expiring supplies. It also allows for facility use changes to be made prior to an increase in disease cases appearing, which increases usage efficiency of facility space and staff.

Overview and Scope of Solution

To address the problem of wasted expenditures in healthcare, I conducted an analysis of the Centers for Disease Control's *Weekly Counts of Deaths by State and Select Causes* datasets and created a report to be used by individual state healthcare systems so that they may redirect physical, human, and financial resources towards reducing the death rate of those specific diseases with the highest growth rates based on statistical significance. By identifying diseases with statistically significant increases in death, healthcare systems will be able to make strategic long-term plans that will: reduce wasted expenditures in the healthcare system, reduce deaths, improve patient satisfaction while reducing patient frustrations, and reduce overall financial costs of healthcare.

The scope of the project is limited to the states and special regions in the United States. The causes of deaths analyzed are limited to natural deaths, septicemia, cancer, diabetes, Alzheimer's, lower respiratory, other respiratory, kidney, abnormal findings, heart disease, brain diseases, Covid-19, influenza and pneumonia. The dates range from 2014 through the 1st quarter of 2021. States' diseases with redacted data at the 33% level were excluded from the statistically significant growth analysis. This removed all the Covid-19 data from that part of the analysis due to it not existed prior to 2019. In addition, no state-to-state or regional comparisons or correlations based on diseases were made due to differences in reporting consistency and population differences between states. Population sizes were used to order the disease and compare data of large states to smaller states.

Project Plan

Overview and Scope of Solution

The goal of the project is to reduce wasted expenditures and deaths in the healthcare system. To accomplish this goal, the objective is to identify diseases with the greatest statistical increase in deaths from year to year for each state or special region in the United States to be used by Healthcare Systems for strategic planning so that they may reduce wasted expenditures and decrease deaths.

The product is the Exploratory Data Analysis code and raw findings detailing the analysis process. The findings of the most significant increases in deaths by disease per state are listed in the Results section below. In addition, this report summarizes the process, methodology and information used in the analysis. To complete the EDA and produce this report, aggregation, cleaning, wrangling, exploration and analysis of the Centers for Disease Control's *Weekly Counts of Deaths by State and Select Causes* datasets are conducted using the R programming language.

Project Planning Methodology and Execution

The project follows an Iterative Waterfall method due to the project objectives and goals being clearly described and the project not being overly complex or large. Iteration is not needed for this report but will be needed for future reports. Cost increases will occur with iteration and customizations requests.

Iteration will be needed as specific report requests are made, responding to formatting and theme customization requests, and targeting the report to specific regional entities. Examples of customizations include refining dates and locations for the reporting, changing fonts and graphical output colors to match the organizations' preferred color libraries.

Milestones and Timelines

Item Description	Start Date	End Date	Estimated Time	Completion Date with Total Hours
Environment setup	7/23/21	7/23/21	0.5 hours	7/23/21 0.5 hrs
Acquire data and initial exploration	7/23/21	7/23/21	1 hour	7/23/21 4 hrs
Data cleaning and wrangling	7/24/21	7/24/21	2 hours	7/24/21 3 hrs
Exploratory data analysis	7/24/21	7/25/21	8 hours	7/25/21 22 hrs
Table/Graphic selection and design	7/25/21	7/25/21	2 hours	7/26/21 4 hrs
Table/Graphics feedback and refinement	7/25/21	7/25/21	1 hour	7/26/21 1 hr
Code refactoring, function creation, optimization	7/25/21	7/25/21	4 hours	7/26/21 2 hrs
Draft analysis report design, layout, outline and creation	7/25/21	7/26/21	3 hours	7/26/21 2 hrs
Analysis report feedback	7/26/21	7/26/21	1 hour	7/27/21 2 hrs
Final analysis report creation	7/26/21	7/26/21	1 hour	7/29/21 6 hrs
Video of code walkthrough and summary of project	7/26/21	7/26/21	3 hours	7/31/21 2 hrs
Project submission	7/26/21	7/26/21	0.1 hours	7/31/21 0.1 hrs
Total projected days and hours		4 Days	26.6 hours	7/31/21 49 hrs

Costs

Item Description	Unit(s)	Unit Cost	Total Cost
Computer with i7+ or equivalent processor, 16Gb+ memory, 512+ SSD (NVME), GPU, Monitor/Mouse/Keyboard (1)	1	$2000 - $3000 [if not available]	$0 [on premises]
R programming language (2)	1	$0	$0
R-Studio (or IDE of choice) (2)	1	$0	$0
Pacman (2)	1	$0	$0
Dplyr (2)	1	$0	$0
Tidyr (2)	1	$0	$0
Knitr (2)	1	$0	$0

Ggplot2 (2)	1	$0	$0
RColorBrewer (2)	1	$0	$0
Psych (2)	1	$0	$0
Data.table (2)	1	$0	$0
Cowplot (2)	1	$0	$0
Hmisc (2)	1	$0	$0
Corrplot (2)	1	$0	$0
Initial Labor Cost (based on $32 hour rounded up to nearest whole hour) (Indeed)	26.6 hours	$32 per hour	$864
Ongoing Labor Cost (Local region reports as requested)	0.5 hours	$32 per hour	$16 per request

(1) if not already available
(2) 3rd party libraries used for project. Other library options are available as preferred

Methodology

Data Collection Process

This report explores the Center for Disease Control's (CDC) Weekly Morbidity and Mortality data from 2014 through the 1st quarter of 2021. The data includes weekly death counts for specific causes in the United States. The data also includes weekly death counts for individual states and specific areas such as New York City, Puerto Rico and Washington, D.C. The data is voluntary and not guaranteed to be reported in a timely or regular basis. Therefore, the most recent data is incomplete and unreliable for analysis and insights. More information on the data set can be found at CDC's website using the following links:

- 2014-2015 MMWR Data Set
- 2020-2021 MMWR Data Set

Data Governance and Privacy

There are not issues or obstacles with the data selection, collection and usage due to the data being open public data. Privacy and security issues are not a concern due to the data sets having already satisfied federal data governance requirements.

Dataset Evaluation: Advantages and Disadvantages

The benefits of using these datasets are that they are in a cleaned format and contain the information needed to solve the organizational need without significant data cleaning or wrangling. No challenges were encountered during collection of the data. Future updates are expected to follow the same procedures and not have any additional time or financial costs. The data is in csv format and no extraction or preparation is needed to load the data. Therefore, the data may be processed and analyzed in multiple freely available no cost or low cost software packages such as Excel, Jupyter Notebooks and R-Studio.

Tools and Reasoning

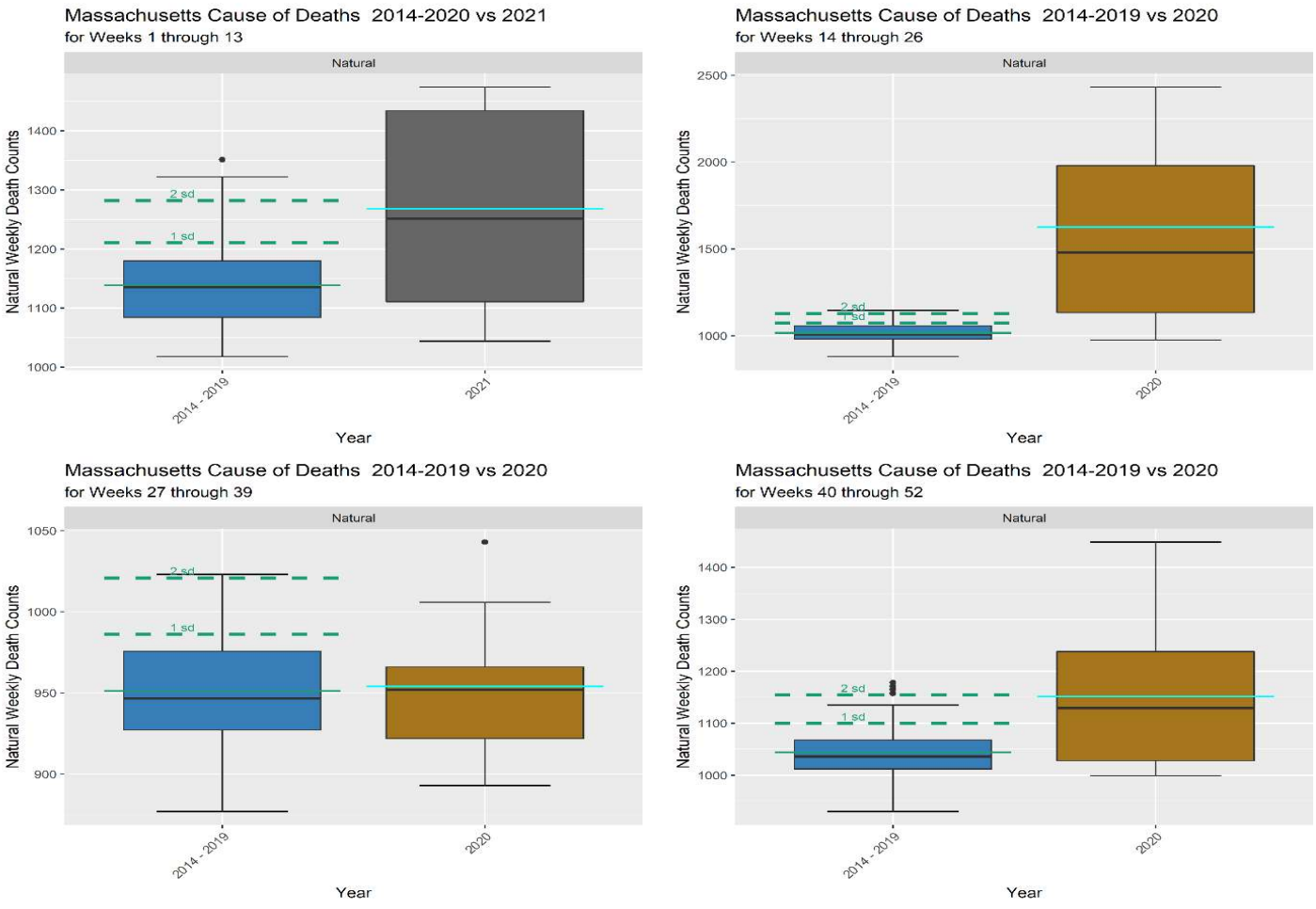
The project analysis is completed using the R programming language, R-Studio® IDE on a Windows 10 machine. The project is not tied to any language or operating system and may be ported to another language as desired. Third party open source libraries used are:

- Pacman for module and package management
- Dplyr for simplifying data selection and filtering statements
- Tidyr for pivoting, organizing and reshaping data

- Knitr for report generation and export to HTML
- Ggplot2 for graphical support and visualization
- RColorBrewer for color palette management and extension
- Psych for correlation and statistical functions and graphics
- Data.table for sql type queries and data structures
- Cowplot for graphical placement and decorators
- Hmisc for adding graphical features and dependencies
- Corrplot for visualizing multi-variate correlation plots

Process, Methods and Techniques

Exploratory data analysis techniques are used to analyze the data. First, multiple datasets are explored to find specific datasets that will work best to solve the organizational needs. Data is gathered and loaded into R-Studio® IDE for structure exploration. Once complete, the data is cleaned by changing the data types, removing unneeded data and changing column names to a more user-friendly formats. Next, using descriptive statistics and visual tools, as in the figure below, analysis is completed on the data. Once the analysis is complete, a summary of finding with recommendations are given based on the objectives of the project, to reduce wasted spending in healthcare systems by planning for and focusing resources on the disease with the highest growth rate. The entire process and code is given in the eda_code.html file for reference and verification as needed.



Statistical Tools

In addition to descriptive statistics used to understand the structure of the data, inferential statistics is also used to evaluate the statistical significance of the disease growth rate and identify the disease in each state with the most significant growth rates as shown in the table below and generate the standard deviation lines in the figure on the previous page. The advantage in using inferential statistics is that it gives a confidence level to judge the significance of the results by. This confidence level is also used in organizing and sorting diseases, states and quarters with the most significant changes in weekly deaths.

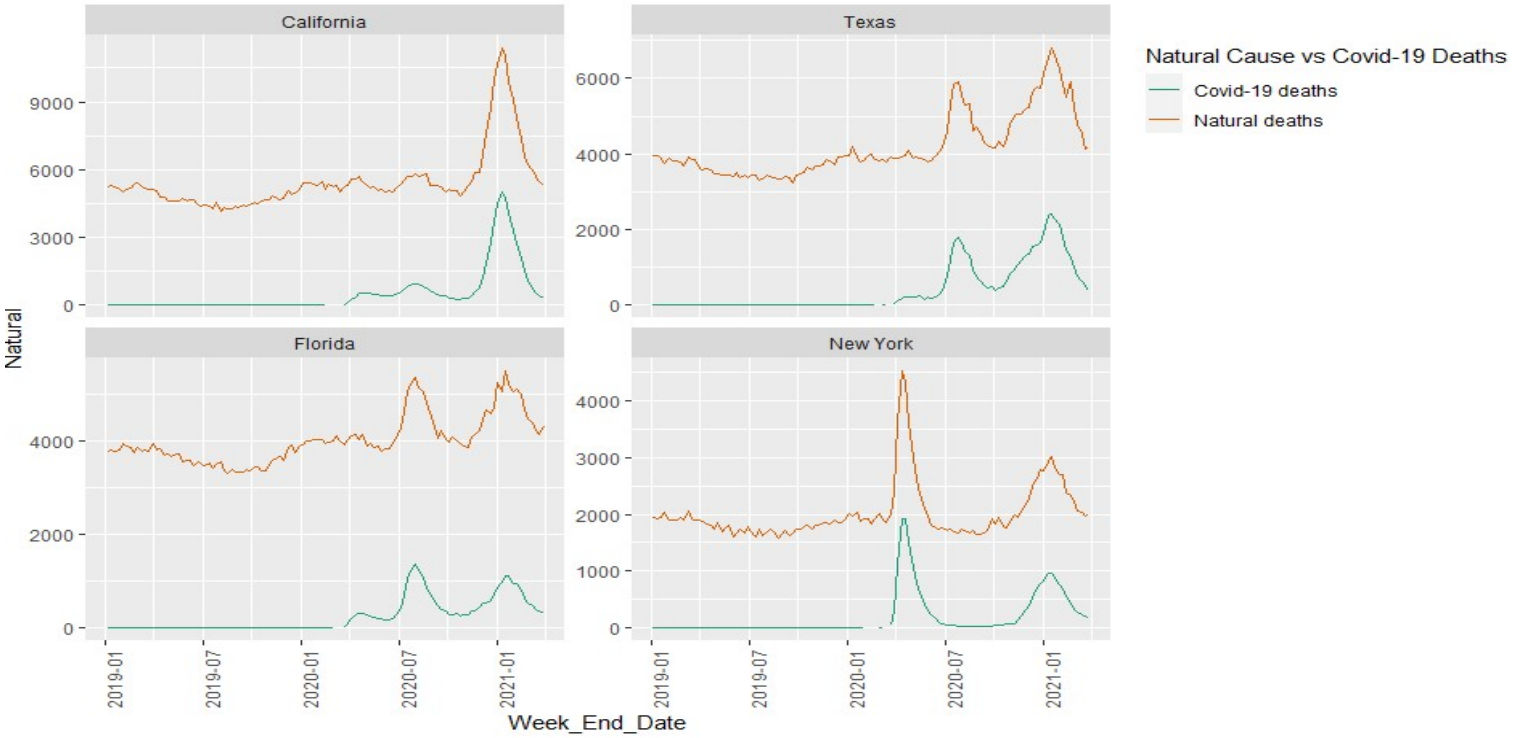
As an example, to create the table below and to determine the states and regions with the most statistical significance, the data for all years before 2020 and 2021 are processed using dplyr's summarize function with qnorm set to 68, 90, 95 and 99 percent significance level. This gives the mean thresholds for the 68, 90, 95 and 99 percent significance levels. The results may then be verified by manually checking state disease samples and through the use of the visual tool in the figure above and later explained in the report.

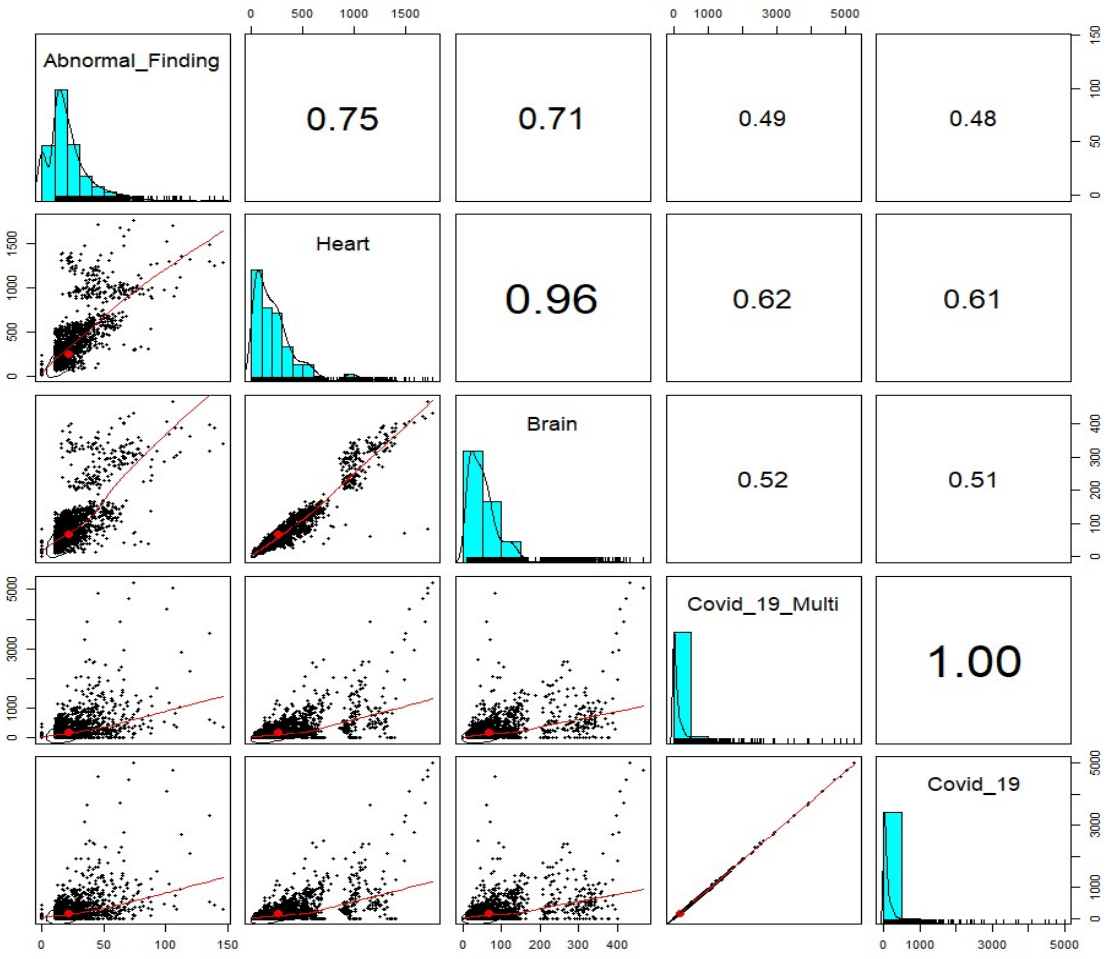
Using the 95th percentile mean (2 standard deviations) as a quotient and the means of the 2020 and 2021 data as the numerator, a ratio is made. This ratio gives a factor that represents how many times the 2020 and 2021 mean is greater than 2 standard deviations or the 95% significance level. If the ratio is 0.5 then the mean is half of 2 standard deviations (1 std dev). If the factor is 1, the mean is at 2 standard deviations and if the factor is 3 it was at 6 standard deviations (3 times 2 std dev). This is done to make comparisons of different disease data easier by normalizing the data to 95%, the target significance level. It was later realized that a ratio of 1 std deviation would make calculating the standard deviation slightly simpler. By then, other functions and graphics were built based on the 2 standard deviation ratio. Therefore, it was not changed. A sample of the resulting output can be seen in the table below with the 90% and 95% confidence ratios given in the last two columns.

Location	Cause of Death	Year	Quarter	90% Confidence Interval Ratio	95% Confidence Interval Ratio
New York City	Natural	2020	Q2	2.79	2.74
New Jersey	Natural	2020	Q2	1.92	1.89
New York City	Heart	2020	Q2	1.83	1.79
New Jersey	Influenza Pneumonia	2020	Q2	1.69	1.59
South Dakota	Natural	2020	Q4	1.55	1.50
Connecticut	Natural	2020	Q2	1.51	1.47
Massachusetts	Natural	2020	Q2	1.49	1.47
New Mexico	Natural	2020	Q4	1.50	1.46
North Dakota	Natural	2020	Q4	1.50	1.46
New York (State)	Natural	2020	Q2	1.48	1.46

Regression analysis is also used to examine correlations of growth rates of similar diseases such as influenza with chronic respiratory disease and possible hidden relationships such as the effect of Covid-19 on natural cause deaths. The advantage of using this technique allows the viewer to find possible relationships between death rates due to misdiagnosis or multiple causes of death. The correlations values imply a possible correlation and not a certain correlation. This is due to the data in the dataset not being from a scientific study. One can only state there is a relationship or possible correlation in the absence of a scientific study.

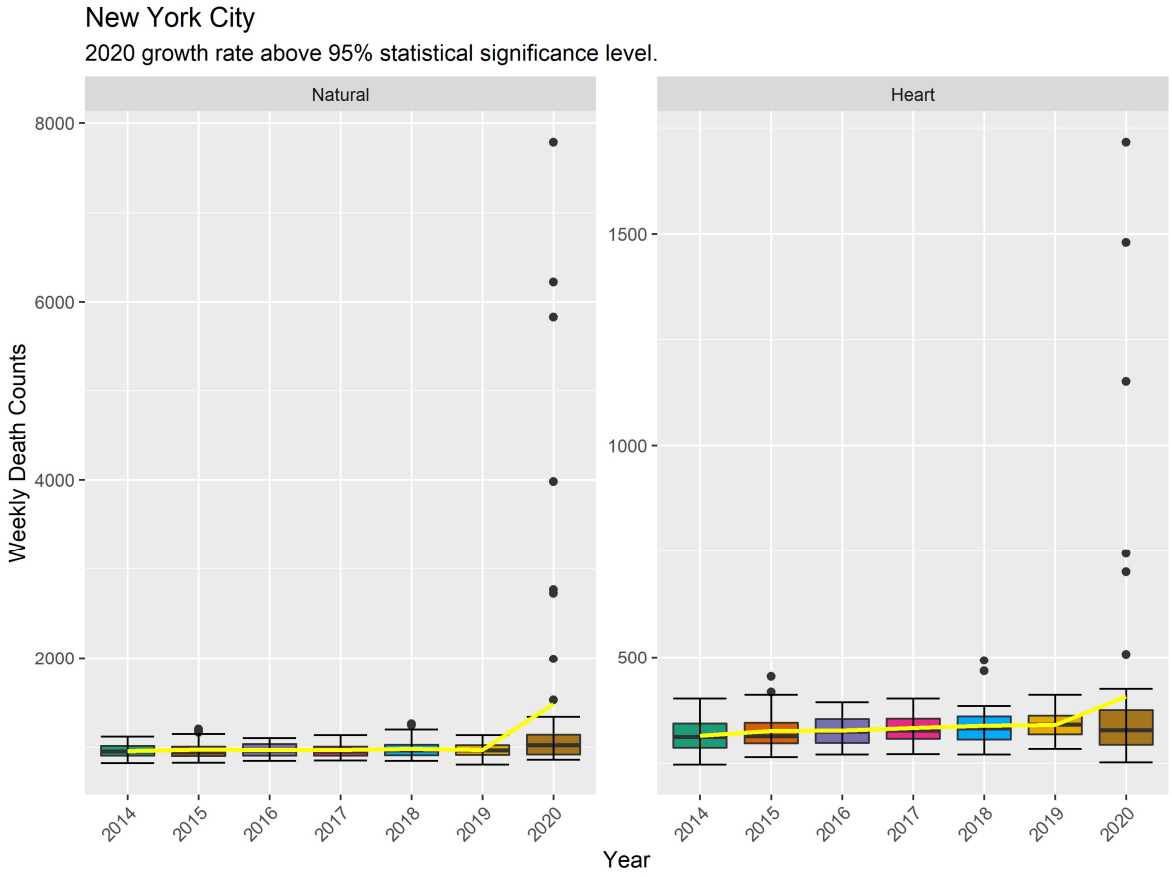
Examples of the possible correlations found between Covid-19 deaths and other disease causes of death are given in the figures on the following page.


California
Texas
Florida
New York
Natural
Week_End_Date
Natural Cause vs Covid-19 Deaths
Covid-19 deaths
Natural deaths

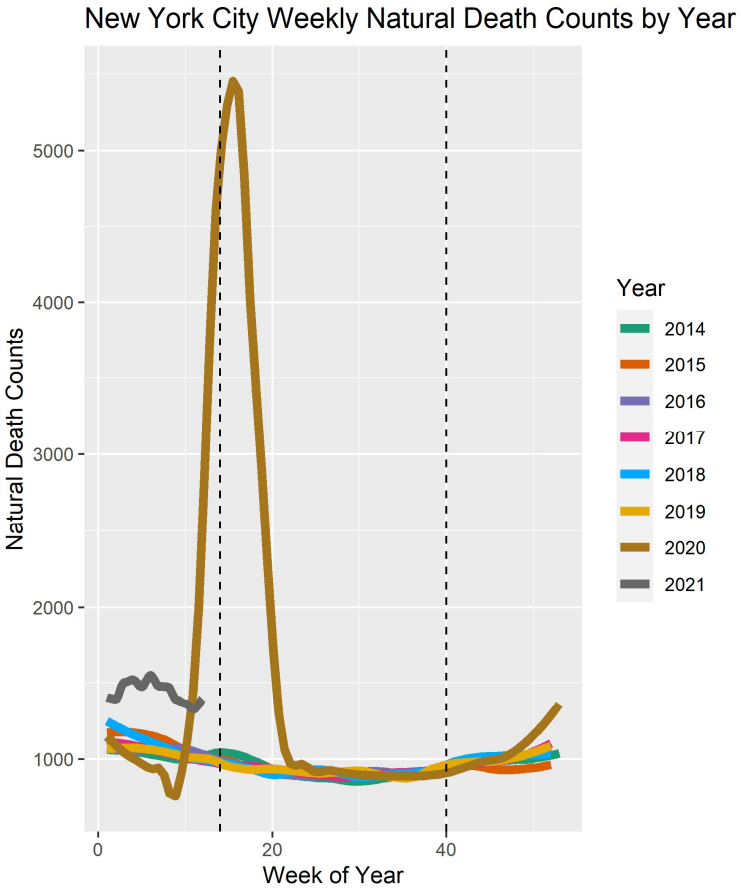

Abnormal_Finding
Heart
Brain
Covid_19_Multi
Covid_19
0.75
0.71
0.49
0.48
0.96
0.62
0.61
0.52
0.51
1.00

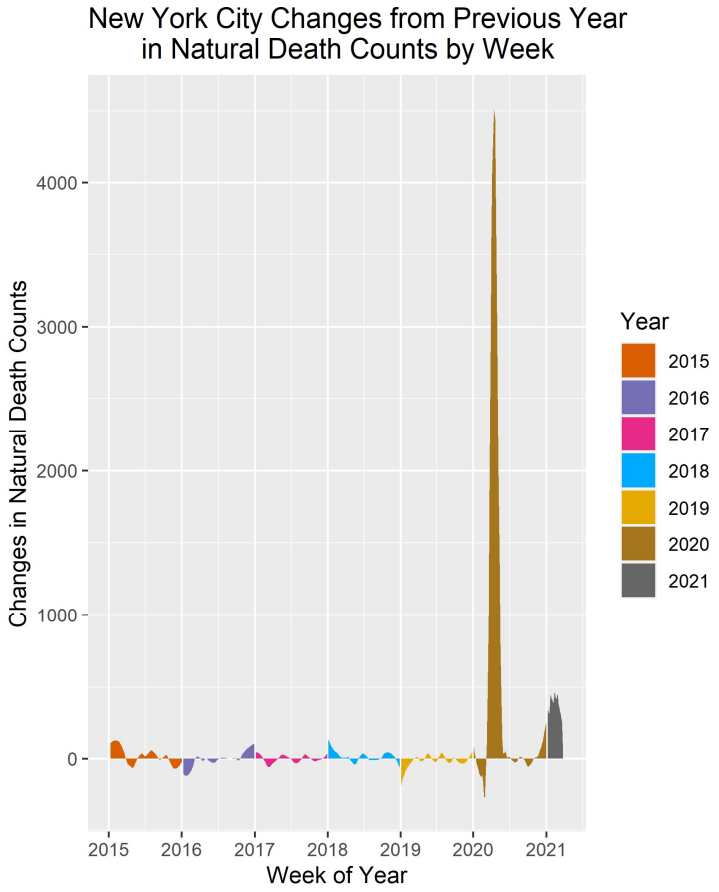
Analysis of the Top 5 States with the Highest Statistically Significant Weekly Death Increases

New York City analysis



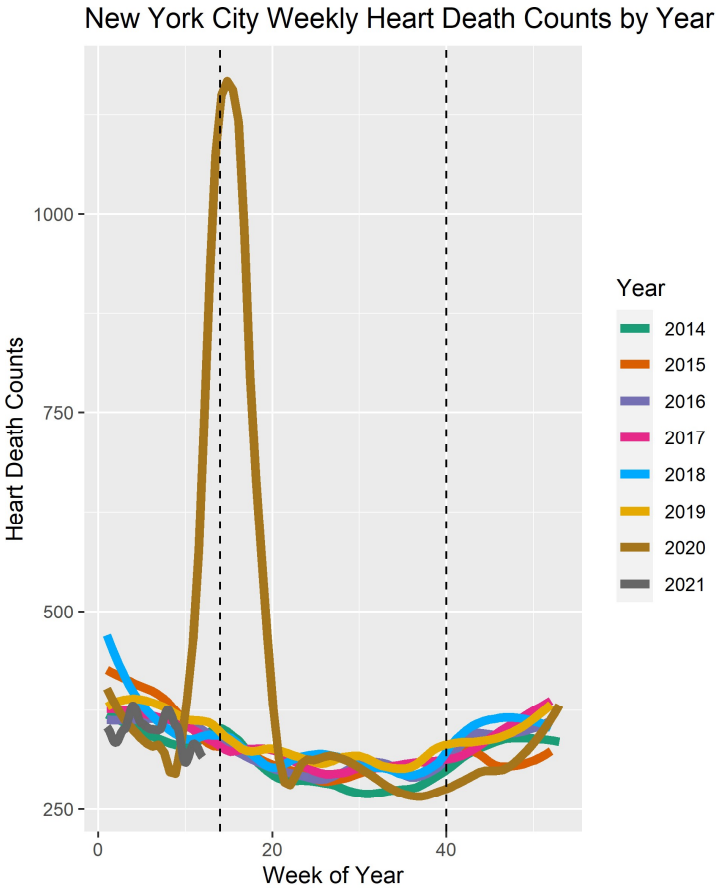
Based on the figures above, there is a significant increase in weekly heart disease and natural cause deaths from 2019 to 2020. Natural cause deaths seem to remain flat before 2019 and heart disease shows a small increase year over year. It is also noted that there are extreme outliers in 2020. These are possibly due to a missed infection of Covid-19 as discussed previously.

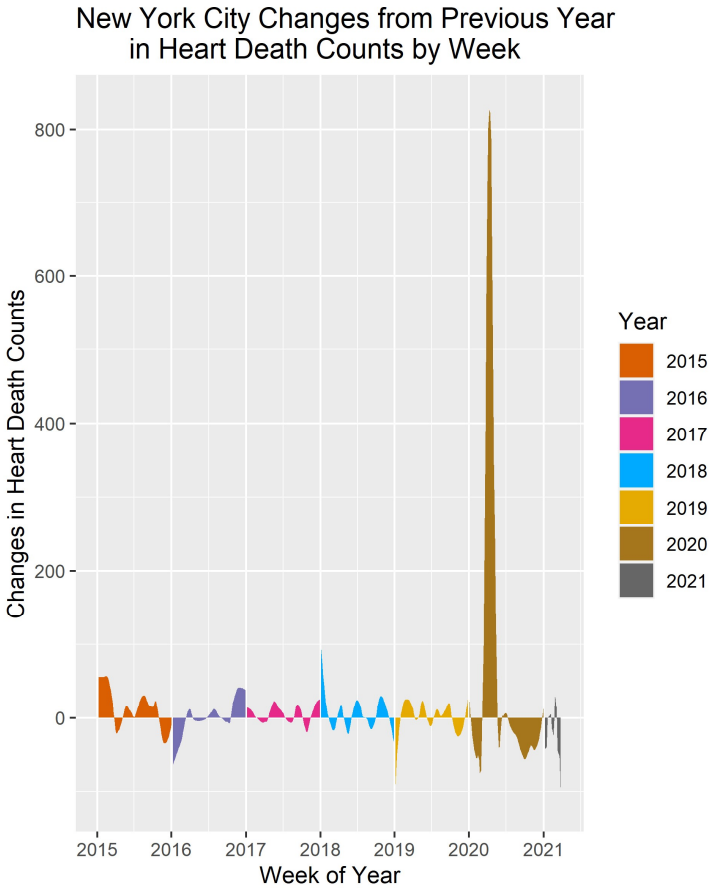




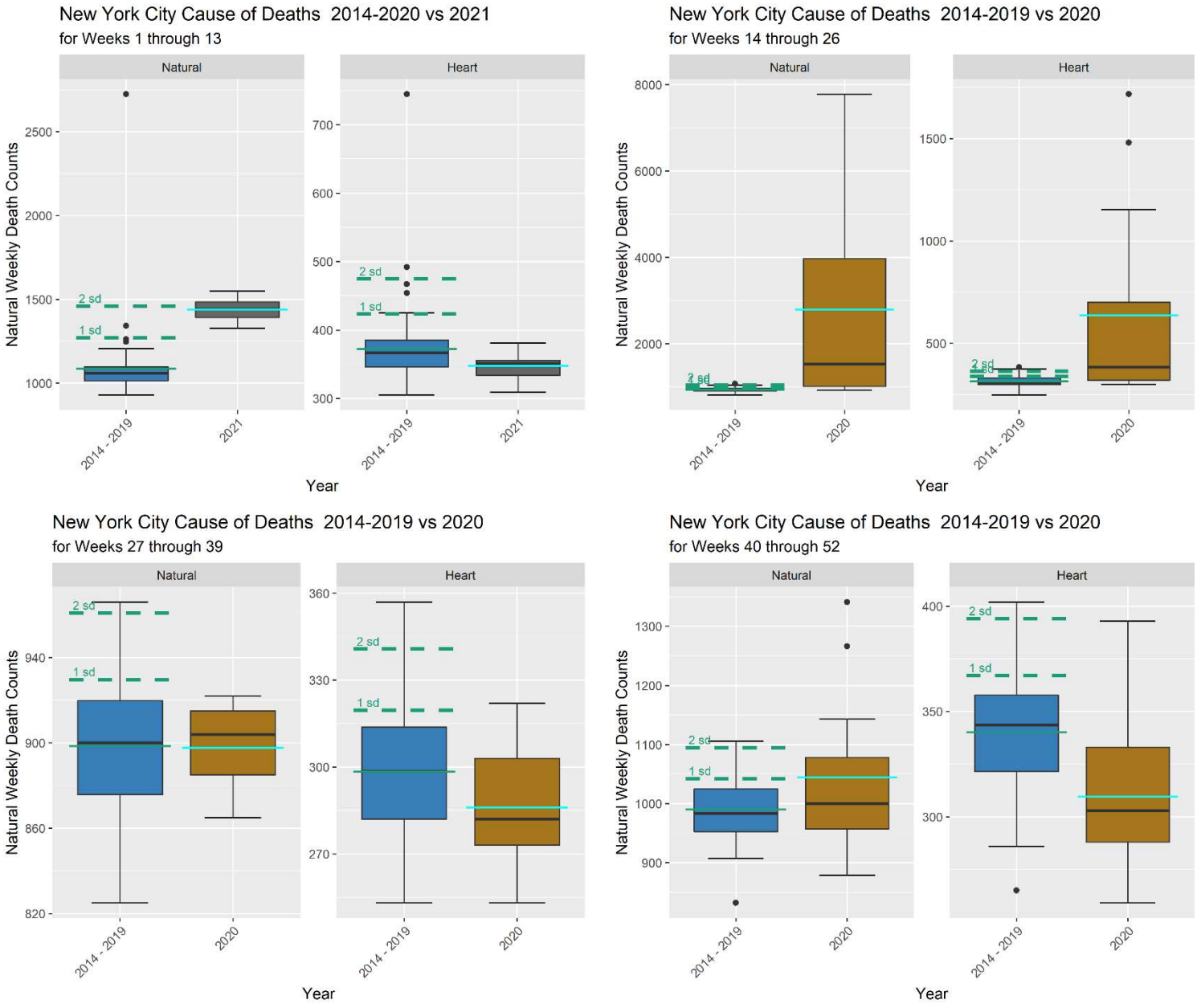
The figures above compare weekly deaths based on each week of the year. The figure on the left shows chronological data that has been smoothed to simplify readability and comparisons. The data in the figure show natural deaths are consistent until 2020 when there is a substantial decrease in weekly deaths, followed by an extreme increase that peaks around 5500 Natural cause deaths just after the first quarter (noted by the first dashed line). The 1st quarter 2021 natural deaths are elevated by approximately 500 additional deaths but are 27% of the 2020 years peak.

The figure on the right shows the changes in weekly deaths from year to year. When the shaded color areas are above 0, the deaths increased for that week when compared to the same week of the previous year. When the shaded areas are below 0, the weekly deaths have decreased for that week compared to the same week of the previous year. Based on the figure on the right, natural causes of deaths have increased compared to years prior to 2020 but are showing a decrease from the extreme peaks of 2020 when comparing 2021 to 2020.





The figure on the left, showing chronological yearly comparisons of weekly heart disease deaths, shows a similar pattern to natural deaths above for 2020. Unlike natural deaths, the figure on the right shows a brief large peak that quickly declines and becomes substantially negative for the remainder of the data. This indicates that the deaths from heart disease have stabilized and were possibly due to comorbidity with Covid-19. Looking back at the left figure, the dark gray 2021 line is below 2019 and even 2018 confirming heart disease deaths have declined. This indicates that the disease is under control. It is interesting that the 2021 data is below the 2019 and 2018 data resulting in possibly 100 less heart disease deaths per week. This is an area that would benefit from future study to determine treatments and strategies that may have decreased heart disease deaths.



The figures above shows the descriptive summary statistics for the most current quarters available in the data (1st quarter of 2021 and 2nd, 3rd and 4th quarters of 2020) compared to the descriptive statistics of the aggregated data of all other years. In addition, there are dashed horizontal lines for the previous years' data indicating the 1st and 2nd standard deviations. These dashed lines identify how significant the current quarter data is when comparing to previous years data. The lighter colored bright-green line on the brown current year's boxplots is the quarter's mean. It should be noted that right skew is common in the data distributions so median (black horizontal line) is a better marker for significant level.

New York City quarterly natural death summary analysis

The data above for weekly natural cause of deaths shows a much greater than 95% statistically significant growth for the 2nd quarter of 2020 followed by a decline in the 3rd quarter and increases in the 4th quarter of 2020 and the 1st quarter of 2021. This shows there is a short term upward trend in the last two quarters with the 1st quarter of 2021 reaching a greater than 95% significance level. These could be related to missed Covid-19 diagnosis which is also likely to have occurred during the 2nd quarter of 2020. Another possibility is the increase in natural deaths are due to more people in the Baby Boomers population bubble reaching geriatric age. It also could be a bi-causality relationship of the two.

Recommendations would be to focus on geriatric populations on the short term, continue monitoring the data to see if it turns into a long-term trend, and conduct small population random testing for the Covid-19 virus to get an accurate estimation of misdiagnosed natural causes of death.

New York City quarterly heart disease death summary analysis

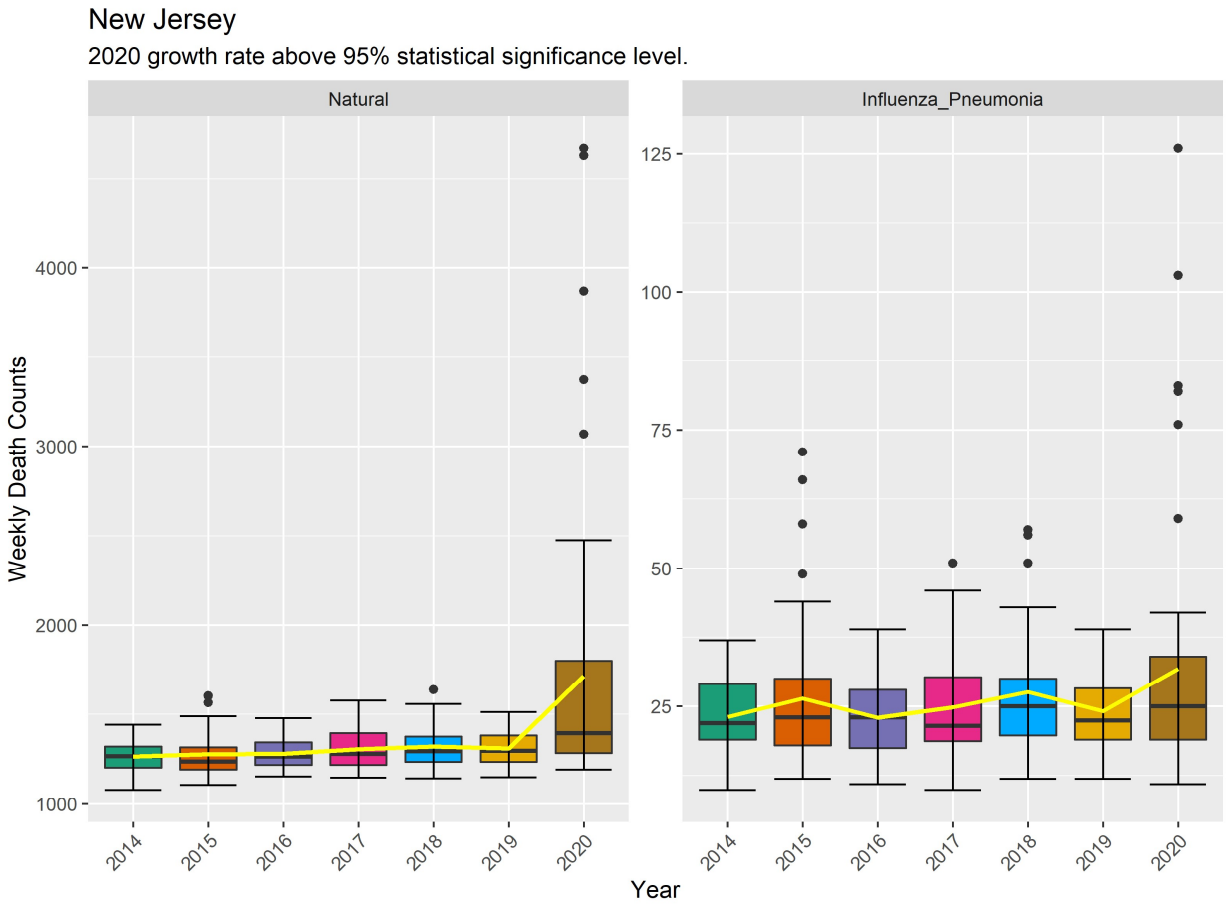
Heart disease deaths were greater than the 95% significance level for the 2nd quarter of 2020 and has since flattened out. In the 4th quarter of 2020, heart disease may have had a significant decrease. Though, this graphic is not designed to detect statistically significant decreases. The recommendation for heart disease is to keep following current practices in place.

New York City summary analysis and recommended course of actions

Based on the findings in the New York City heart disease and natural cause of death categories, I recommend that New York City healthcare systems:

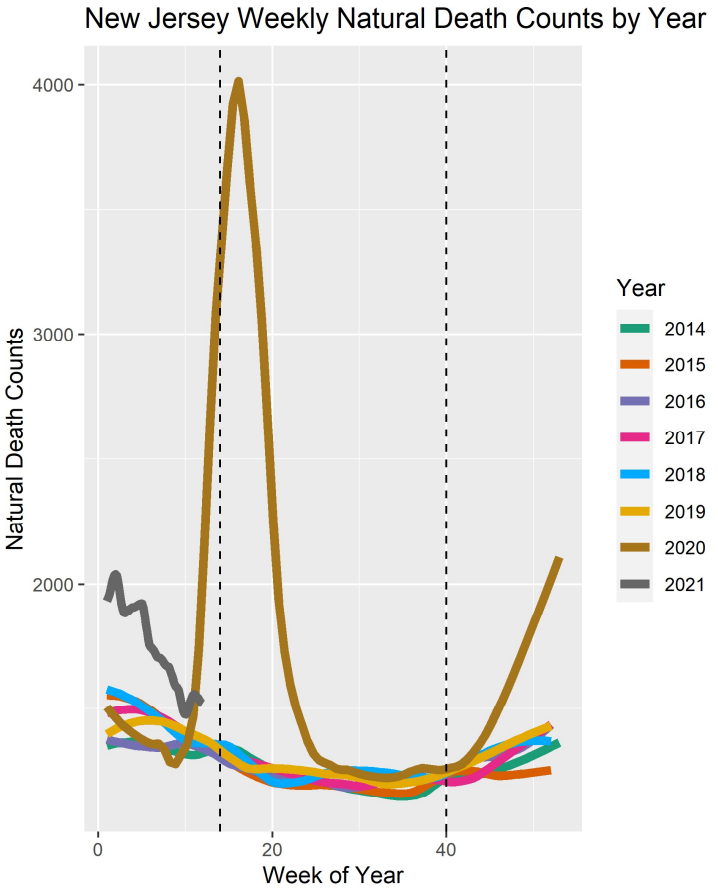
1. Work with their geriatric (elderly) populations and caregivers to ensure that their physical, nutritional, medical, emotional, and social needs are being met. This older population might be used to going it alone and may not reach out for help.
2. Focus short term resources on reducing natural causes of death and continue monitoring for long term trends.
3. For heart disease deaths, current practices are decreasing weekly deaths. Continue following current best practices and monitor the data for any changes.

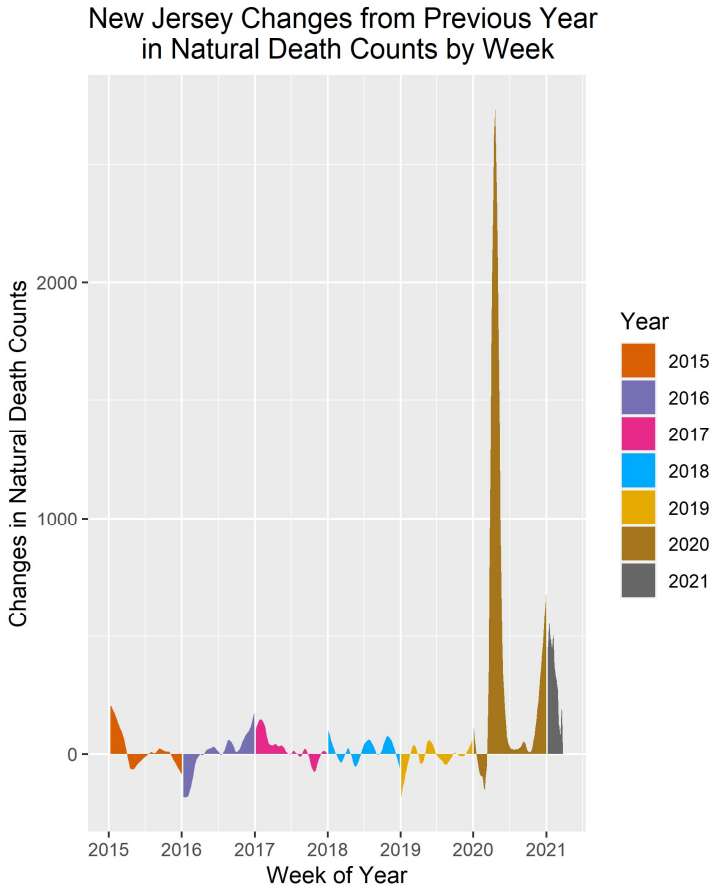
New Jersey analysis



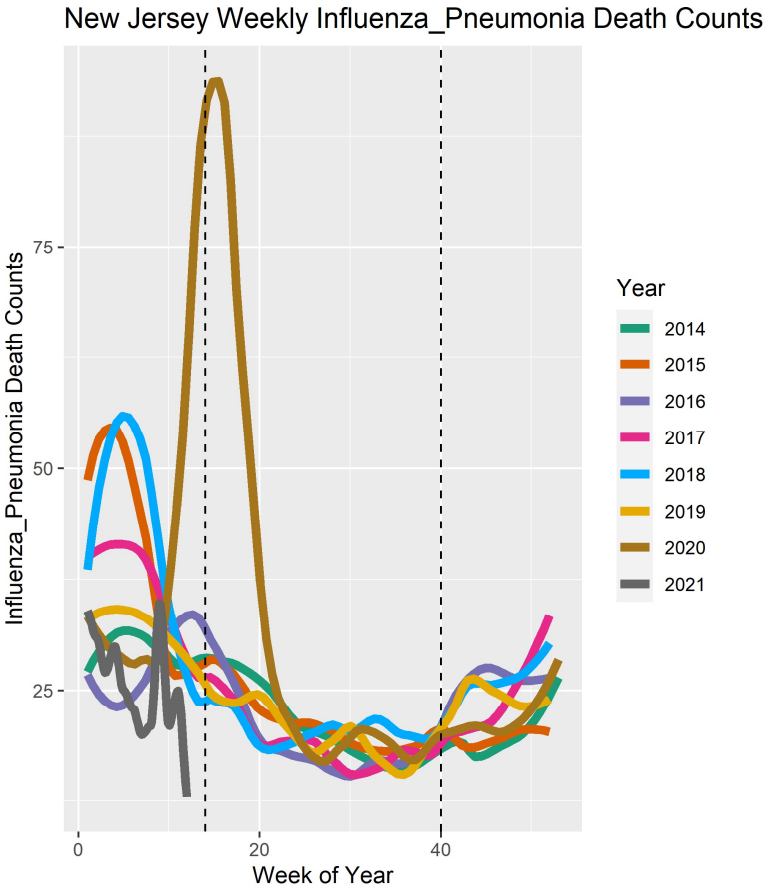
Based on the figures above, natural causes of death are increasing year-over-year showing a long-term trend. There is a greater increase from 2019 to 2020 and several outliers exist in the 2020 data. The increase and outliers are likely due to the same reasons as given in the New York City data, Covid-19 misdiagnosis or aging population bubble.

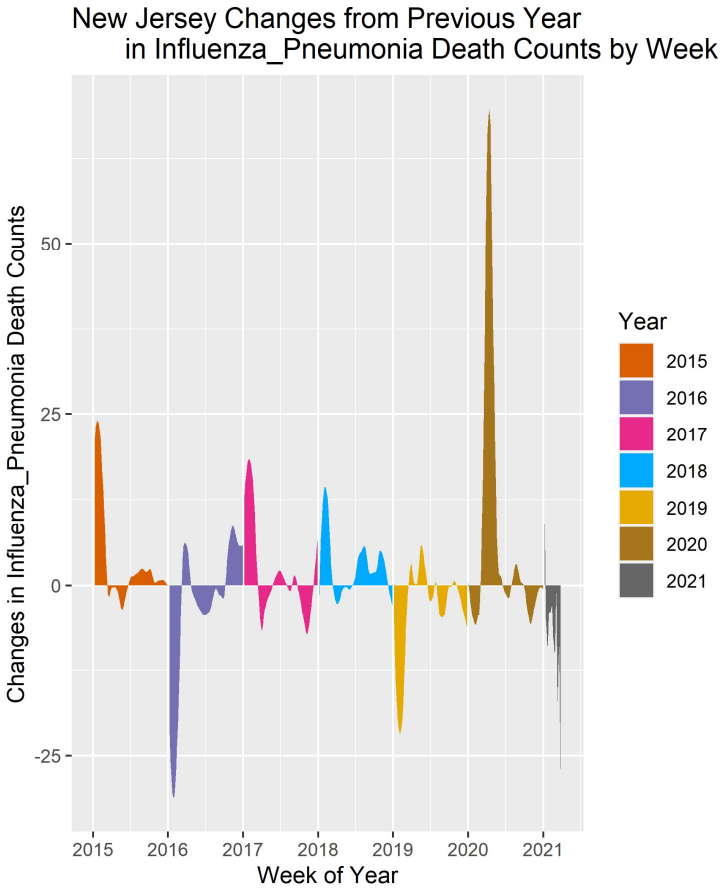
The influenza and pneumonia data remains consistent until 2020 when there is a greater growth rate.



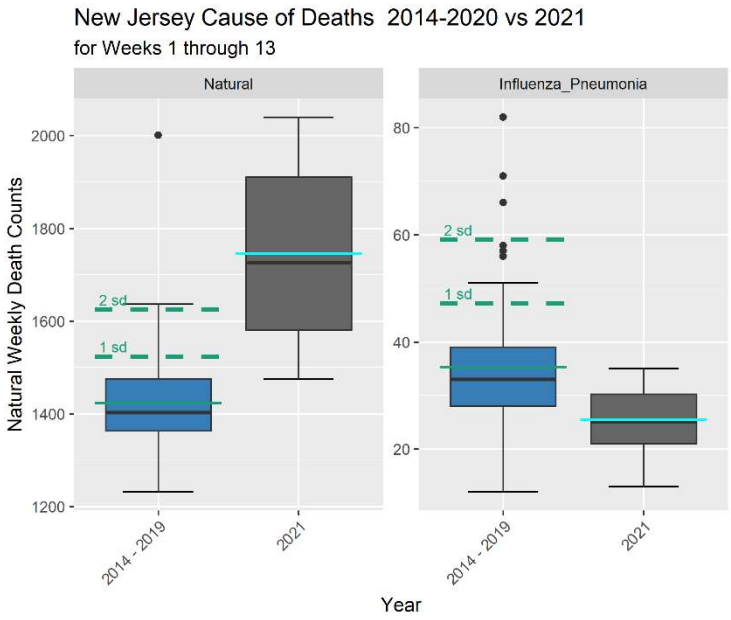


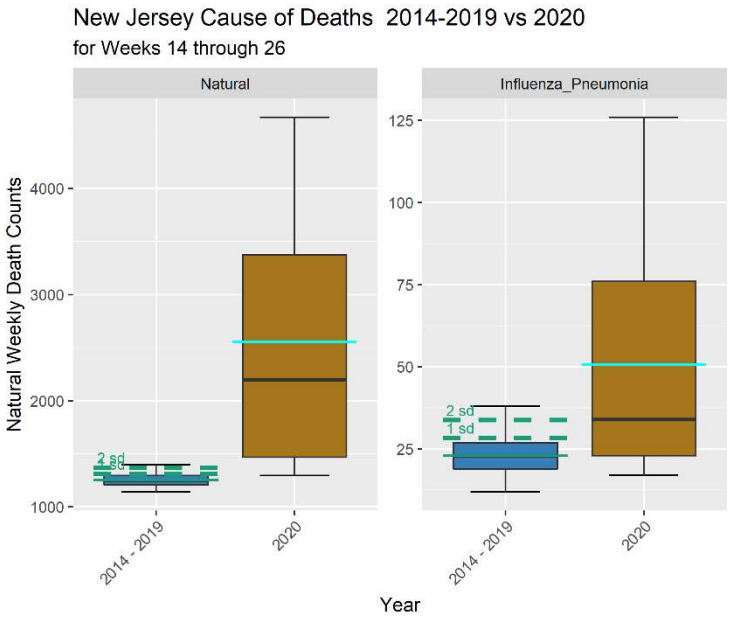
The left figure for natural cause weekly deaths shows a similar pattern to the New York city data with more a more pronounced increase in the 4th quarter of 2020 and a steeper decrease in the 1st quarter of 2021. The 1st quarter 2020 data declines from 50% to 25% of the 2020 peak and is still significantly higher than the 1st quarter of 2020. The significance of this increase is especially noticeable in the right figure where year over year changes in weekly deaths have remained positive for most of 2020 and 2021.

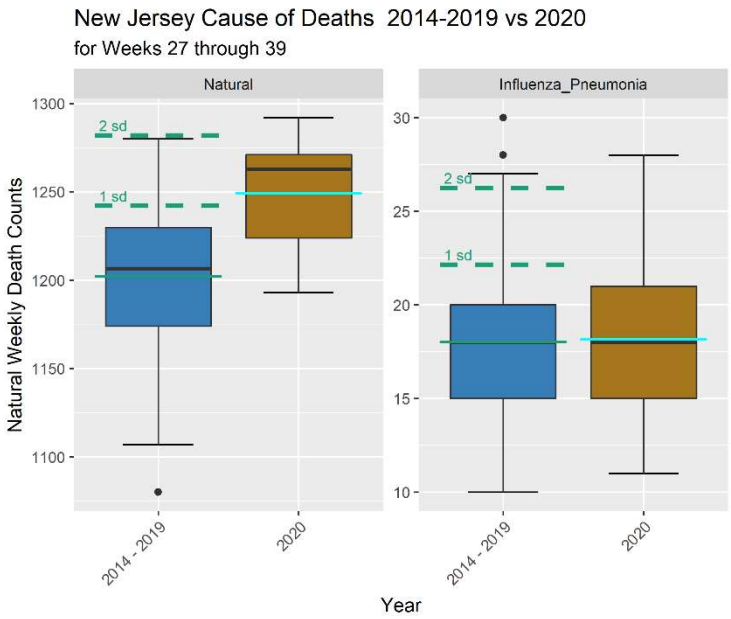


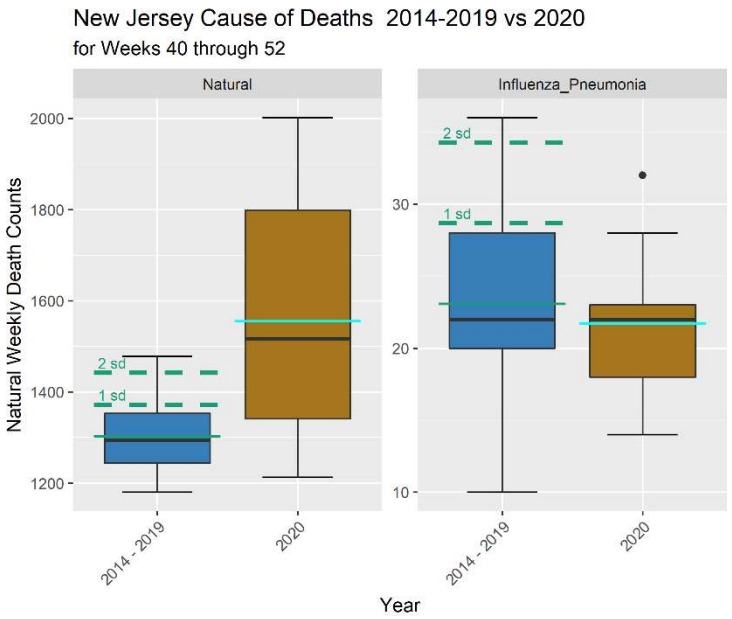


Based on the data in the left figure above, New Jersey had high influenza weekly death counts in the 1st quarters of 2014, 2015, 2017, 2018 and 2019 and then smaller, more consistent, influenza and pneumonia deaths for the remaining quarters. The figure on the right shows alternating increases in influenza and pneumonia weekly deaths, when comparing year over year growth rates, followed by decreases. The growth rates look consistent except for the high positive peak in the 1st and 2nd quarters of 2020. The general patterns seen do not indicate year-over-year long term trends. Though, there is definite long term trend when comparing a year's 1st quarter to the previous year's 1st quarter data. Recommendations would be to expend resources toward reducing 1st quarter deaths from influenza.









New Jersey Quarterly natural death summary analysis

The data for weekly natural cause of deaths above shows significantly greater than 2 standard deviations increases for the 2nd and 4th quarters of 2020 and the 1st quarter of 2021. These are also observed in the line and area graphs discussed earlier though the significance of the increases are clearer in the box plots above.

New Jersey Quarterly influenza and pneumonia death summary analysis

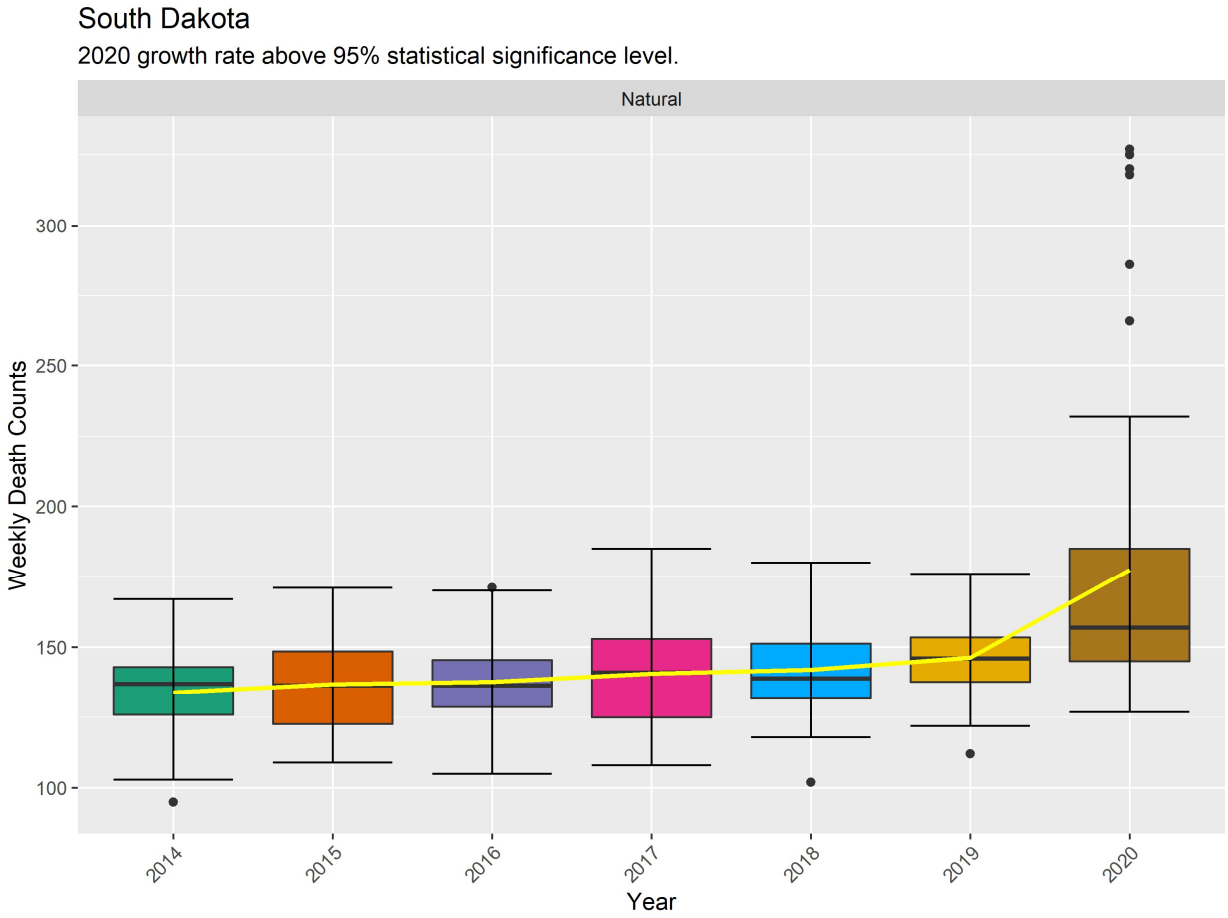
Influenza and pneumonia weekly deaths have a greater than 95% statistically significant increase in the 2nd quarter of 2020. The remaining quarters show decreases in weekly deaths. The 1st quarter 2021 data shows a decrease though there is a notable long-term trend of 1st quarter increases in influenza and pneumonia deaths for other years. The 2021 decrease in weekly deaths shows the effect of extra precautions and social distancing has had on deaths. Therefore, resources should be targeted towards reducing 1st quarter influenza and flu deaths using methods that have shown to be effective.

New Jersey summary analysis and recommended course of actions

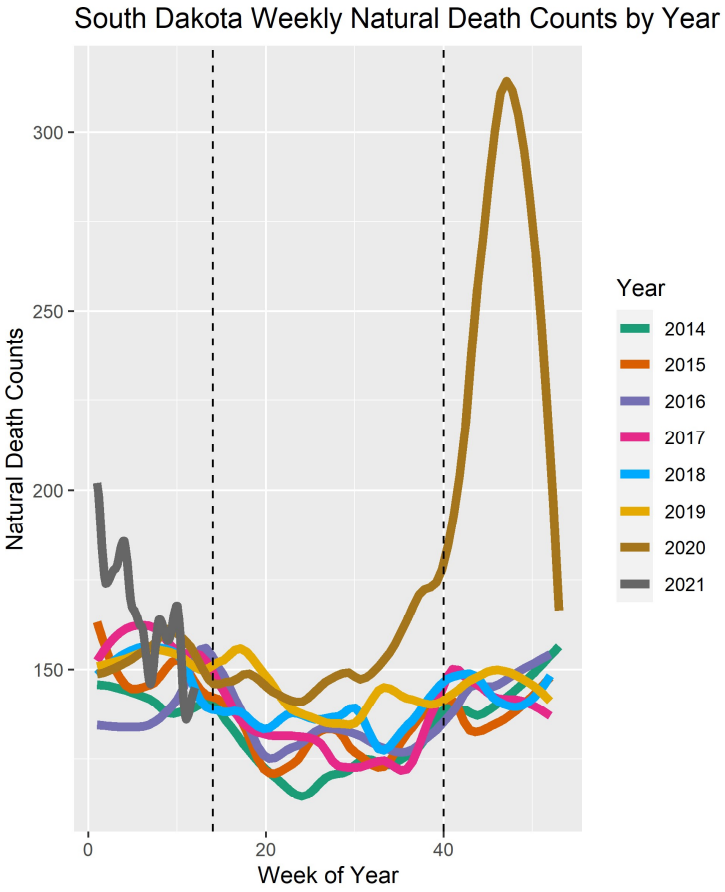
Based on the findings in the New Jersey natural, influenza and pneumonia cause of death categories, I recommend that New Jersey healthcare systems:

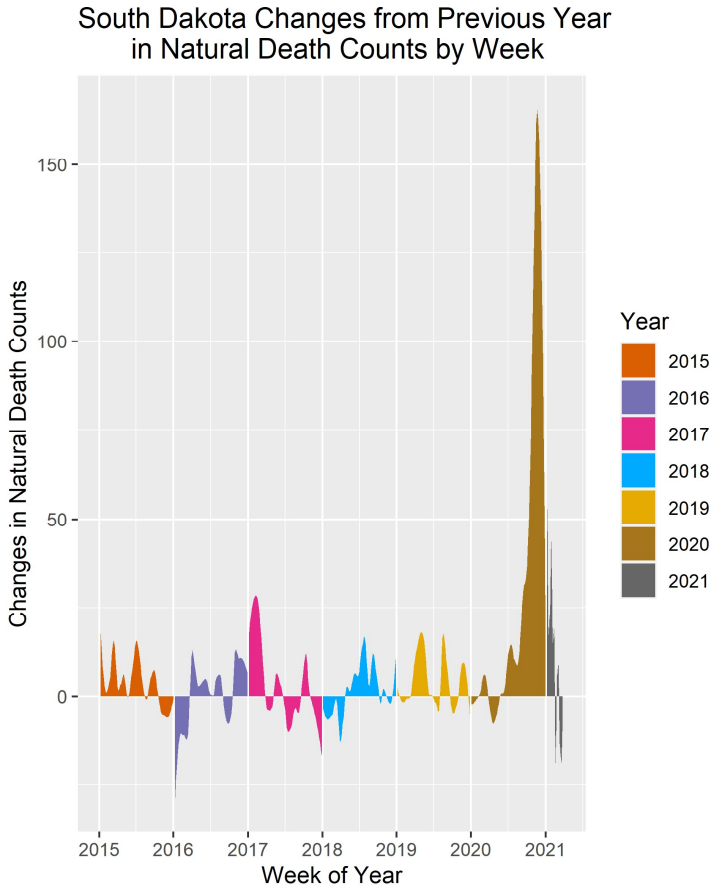
1. Work with their geriatric (elderly) populations and caregivers to ensure that their physical, nutritional, medical, emotional and social needs are being met. This older population might be used to going it alone and may not reach out for help.
2. Focus long term resources on reducing natural causes of death and continue monitoring for flattening of the long-term trend.
3. Focus short term resources towards reducing influenza and pneumonia deaths in the first quarters of the year.
4. Identify procedures and protocols that were effective during social distancing and promote the non-invasive ones that will be easily accepted by the public. A possible example is signage and reminders to wash hands frequently.

South Dakota analysis

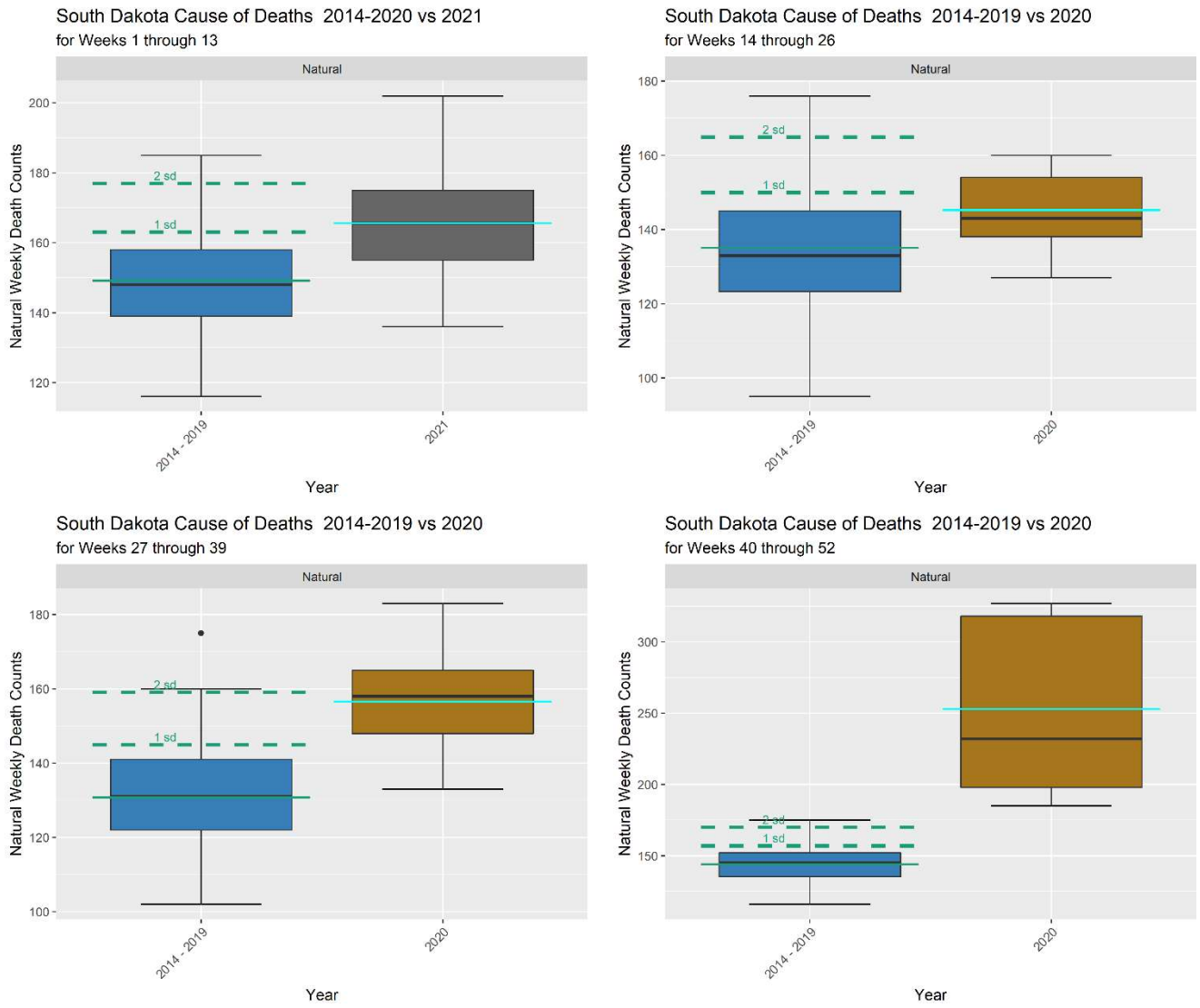


Based on the figure above, natural causes of death are increasing year-over-year showing a long-term trend. There is the same significantly greater increase from 2019 to 2020 seen previously including several outliers in the 2020 data. This is likely due to the same reasons as given in the New York City and New Jersey data: Covid-19 misdiagnosis or aging population bubble. Increases in natural causes of death seems to be a nationwide issue and may benefit from national level resources being applied to reduce the number of weekly deaths.





The data represented in these graphs are slightly different than those previously seen. This indicates a lag or some other anomaly in weekly deaths. The peak for natural causes of death occurs in the 4th quarter or fall of 2020. In other states, the peak occurred at the end of the 1st quarter and beginning of the 2nd quarter of 2020. The figure on the left shows that deaths in this category are declining for the 1st quarter of 2021. The 2021 1st quarter has a notable point around week 11 where weekly deaths are lower than all the other years' week 11 data. This can be seen in the right figure, where the 2021 growth rate data is negative. This indicates that South Dakota has implemented policies and strategies to reduce natural causes of death. The recommendation is that South Dakota healthcare systems disseminate policies and strategies used to reduce natural causes of deaths to other states for evaluation and implementation in other states. Other recommendations are the same as the other states with the addition that resources on a national level might be needed.



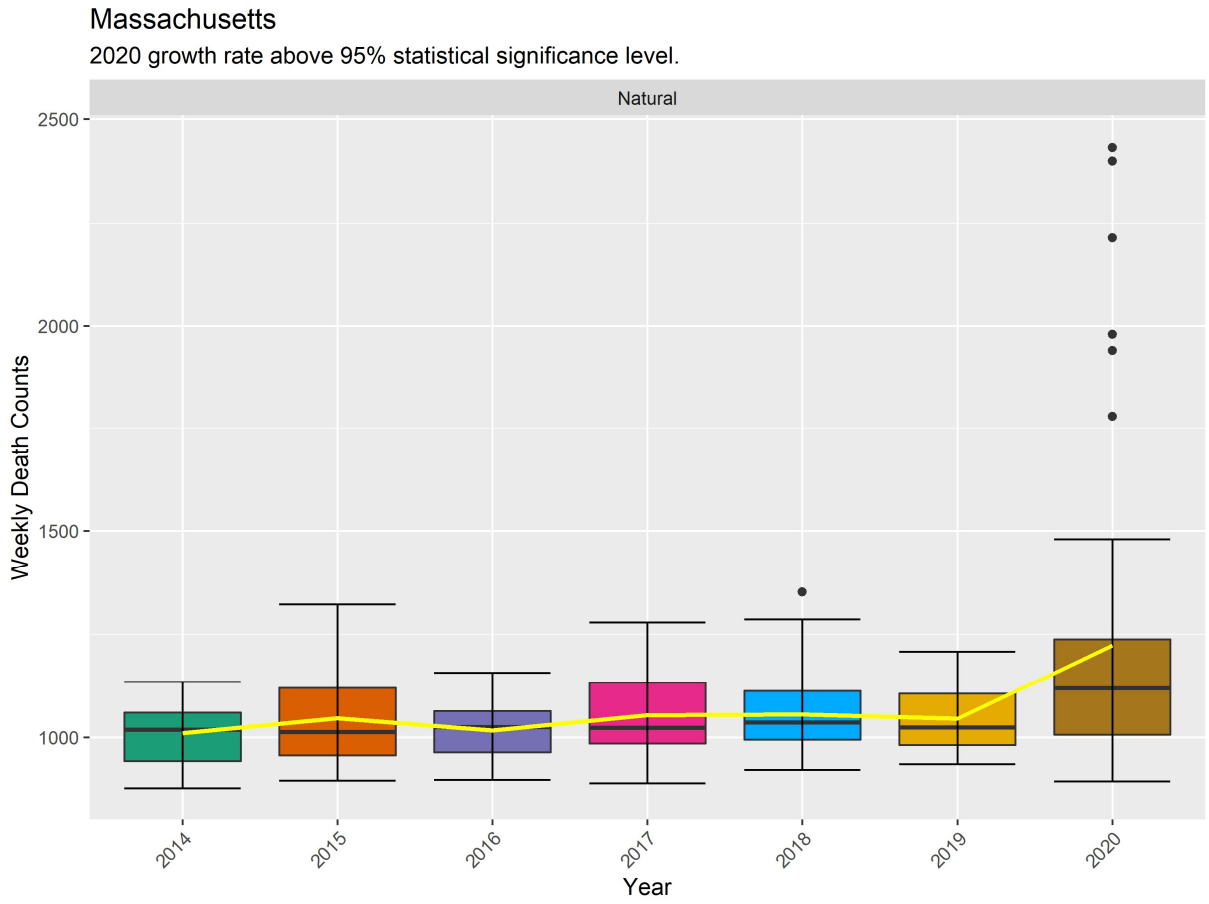
The figures above show the significance of the 4th quarter natural cause weekly deaths. It is much greater than 2 standard deviations with all the first quartile well above 2 standard deviations.

South Dakota summary analysis and recommended course of actions

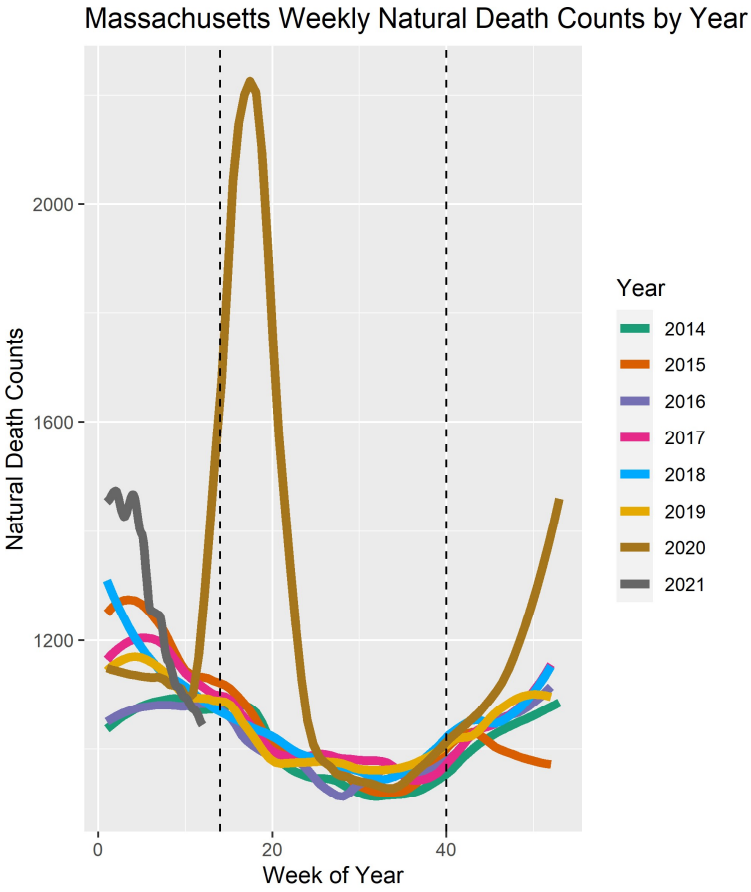
Based on the findings in the South Dakota natural causes of weekly death categories, I recommend that South Dakota healthcare systems follow the same recommendations given earlier and listed below. The only addition, which I recommend for previous states, is to seek national resources and collaboration to reduce deaths in the natural death category

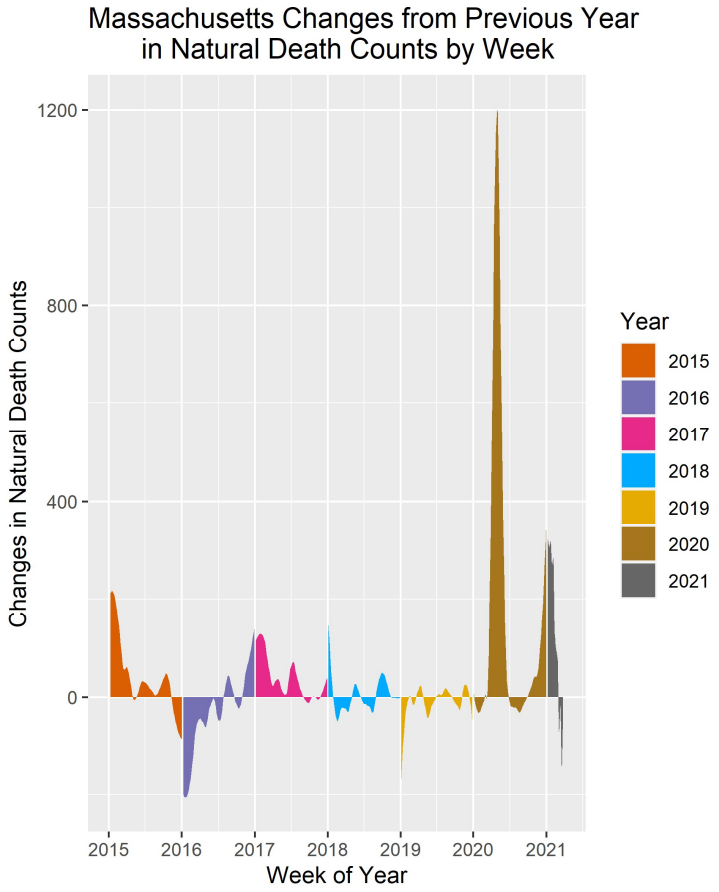
1. Work with their geriatric (elderly) populations and caregivers to ensure that their physical, nutritional, medical, emotional, and social needs are being met. This older population might be used to going it alone and may not reach out for help.
2. Focus long term resources on reducing natural causes of death and continue monitoring for flattening of the long term-trend.
3. Seek assistance and resources at the national level.
4. Collaborate with other states and regions to determine best practices and assess if they can implemented South Dakota's policies and strategies to reduce natural causes of deaths. This should be followed with monitoring the data to see the effect on the long term-trends.

Massachusetts analysis

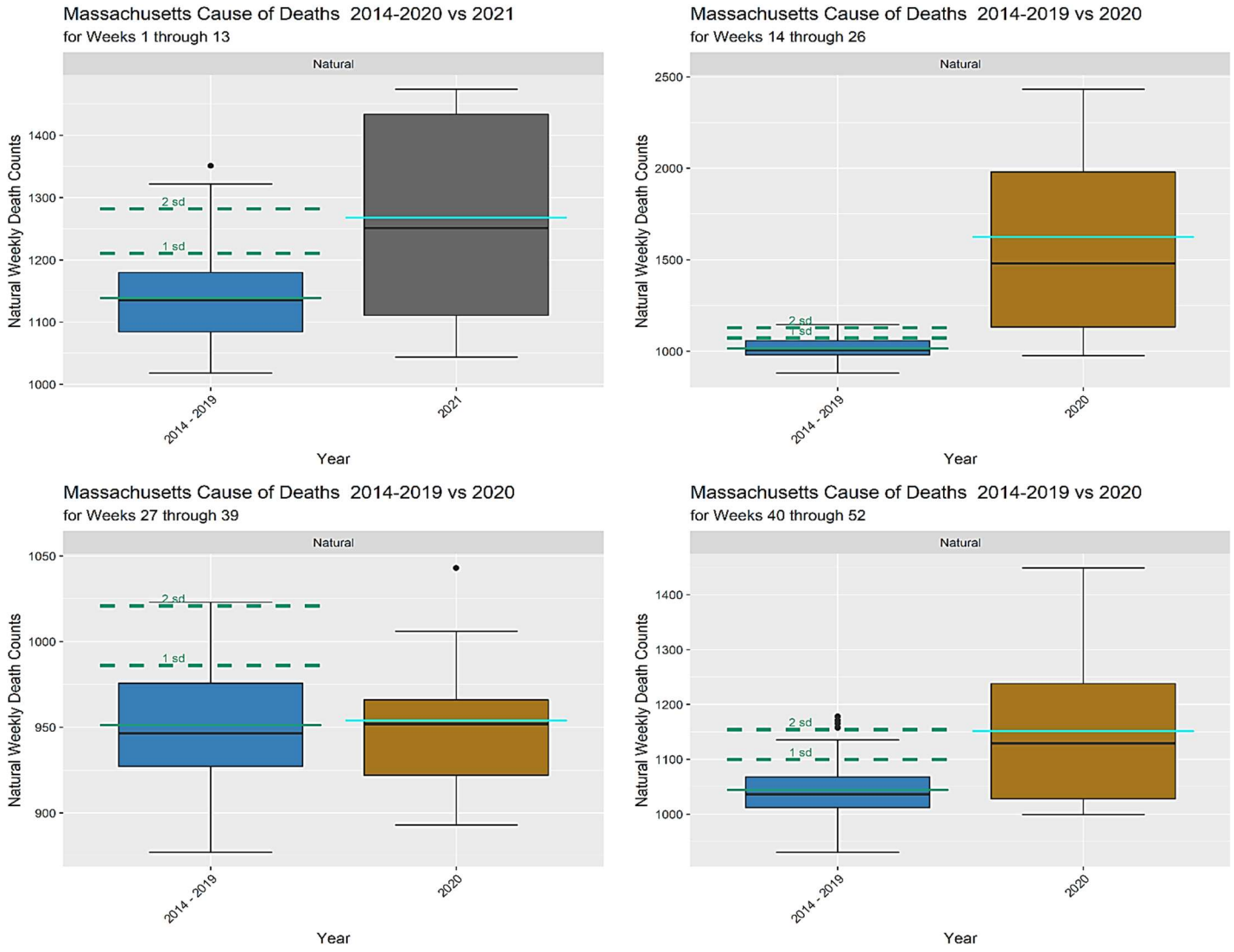


The data above show less of year-over-year increase in the natural cause of weekly deaths and may be considered constant except for the familiar increase from 2019 to 2020. This indicates a short-term growth trend and not a long-term trend.





The figure on the left is similar to the New Jersey data but on a smaller scale. Natural cause weekly deaths increased significantly in the 2nd and 4th quarters of 2020 with the 4th quarter peak continuing into the 1st quarter of 2021. The figure on the right shows decreasing variations in the weekly death rates from 2015 through 2019. After 2019, the deaths increase with the exception of two small periods in the 1st and 3rd quarters of 2020. Similar to South Dakota, weekly natural deaths have decreased below all other years at the end of the 1st quarter of 2021. Massachusetts' healthcare systems should compare policies and strategies that are similar to South Dakota's to identify and isolate the two states' overlapping policies that may be reducing natural causes of deaths. Once identified, these policies should be implemented in other states to determine the possible benefits of a nationwide implementation.



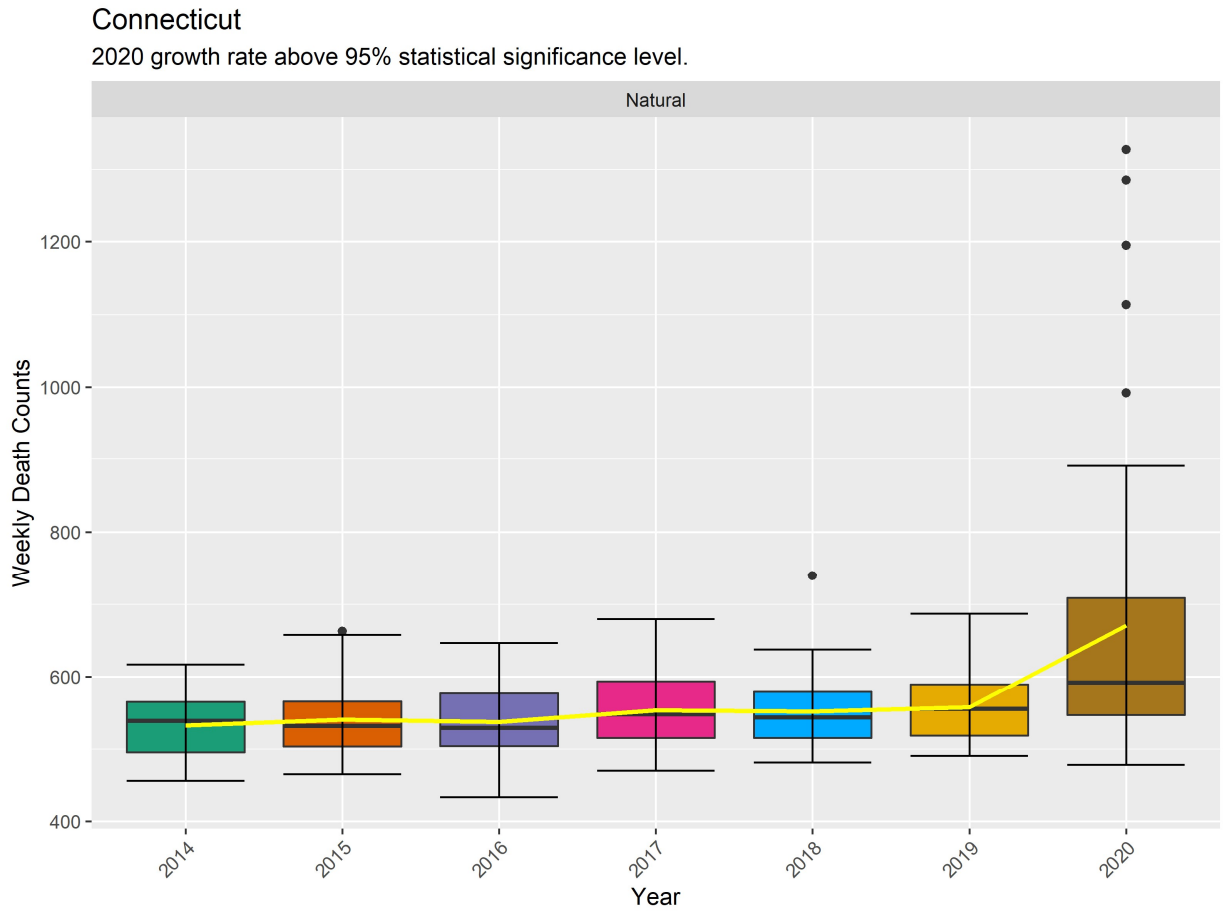
Based on the data shown in the figures above, it is noted that only the 2nd quarter of 2020 has an extreme significance level for natural causes of weekly deaths. The 4th quarter of 2020 and 1st quarter deaths of 2021 are statistically significant but not as extreme as seen in the New Jersey state data.

Massachusetts summary analysis and recommended course of actions

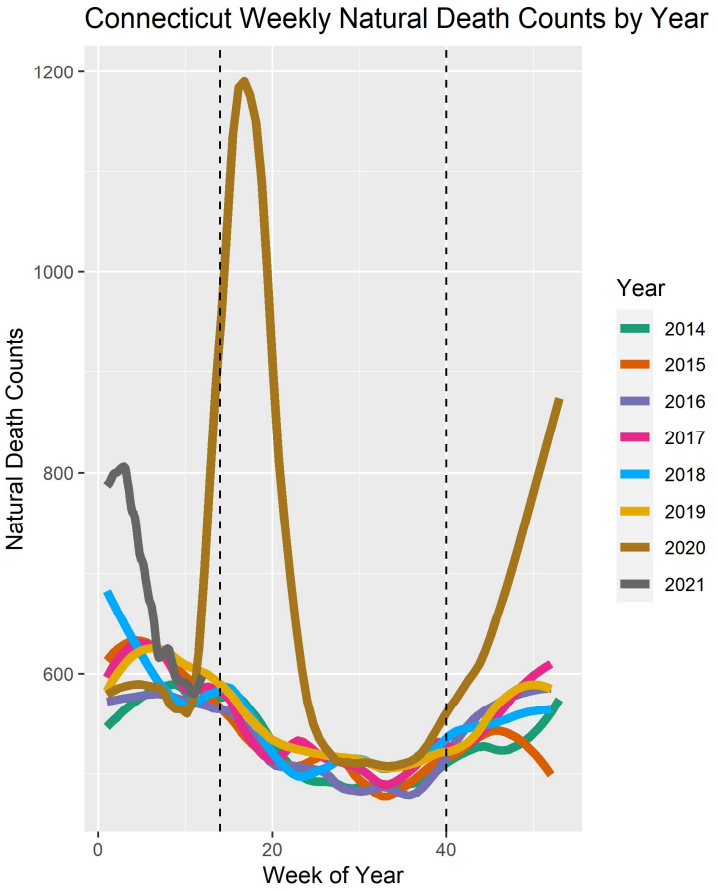
Based on the findings in the Massachusetts natural causes of weekly death categories, Recommendations for Massachusetts' healthcare systems follow the same recommendations given earlier and are listed below.

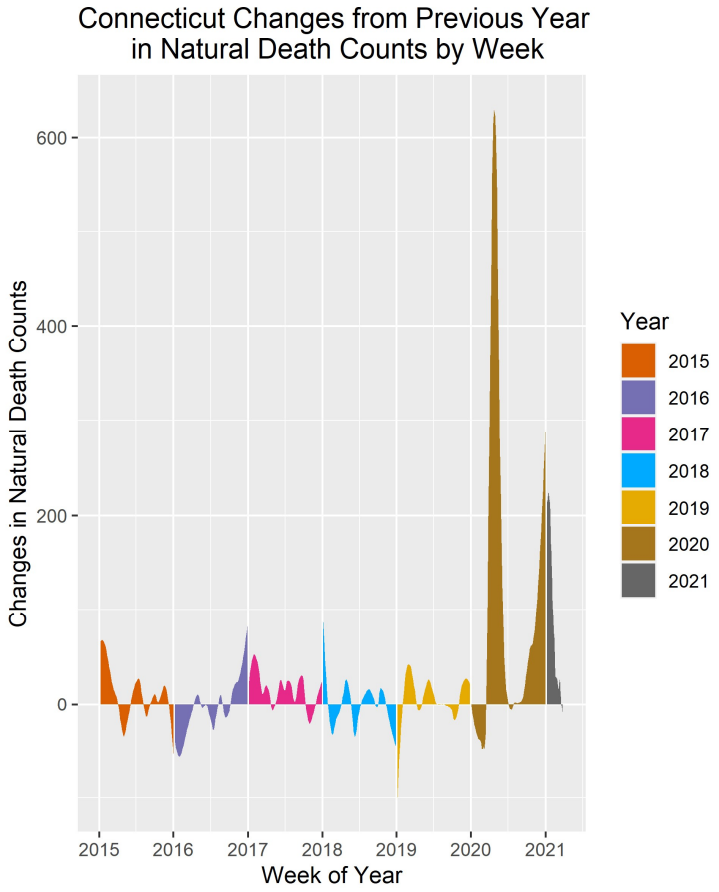
1. Work with their geriatric (elderly) populations and caregivers to ensure their physical, nutritional, medical, emotional, and social needs are being met. This older population might be used to going it alone and may not reach out for help.
2. Focus long term resources on reducing natural causes of death and continue monitoring long-term trends.
3. Seek assistance and resources at the national level.
4. Collaborate with South Dakota to determine overlapping policies that may be reducing natural causes of death. Once identified, disseminate the polices to other states and regions so that they may assess if the policies have similar effects on their weekly natural deaths. If the study is found to cause a statistically significant reduction in weekly natural causes of deaths, the policies and strategies should be broadly implemented. Broad implementation should be followed with monitoring of future data for a reduction in natural causes of death and the effect on the national long-term trend.

Connecticut analysis

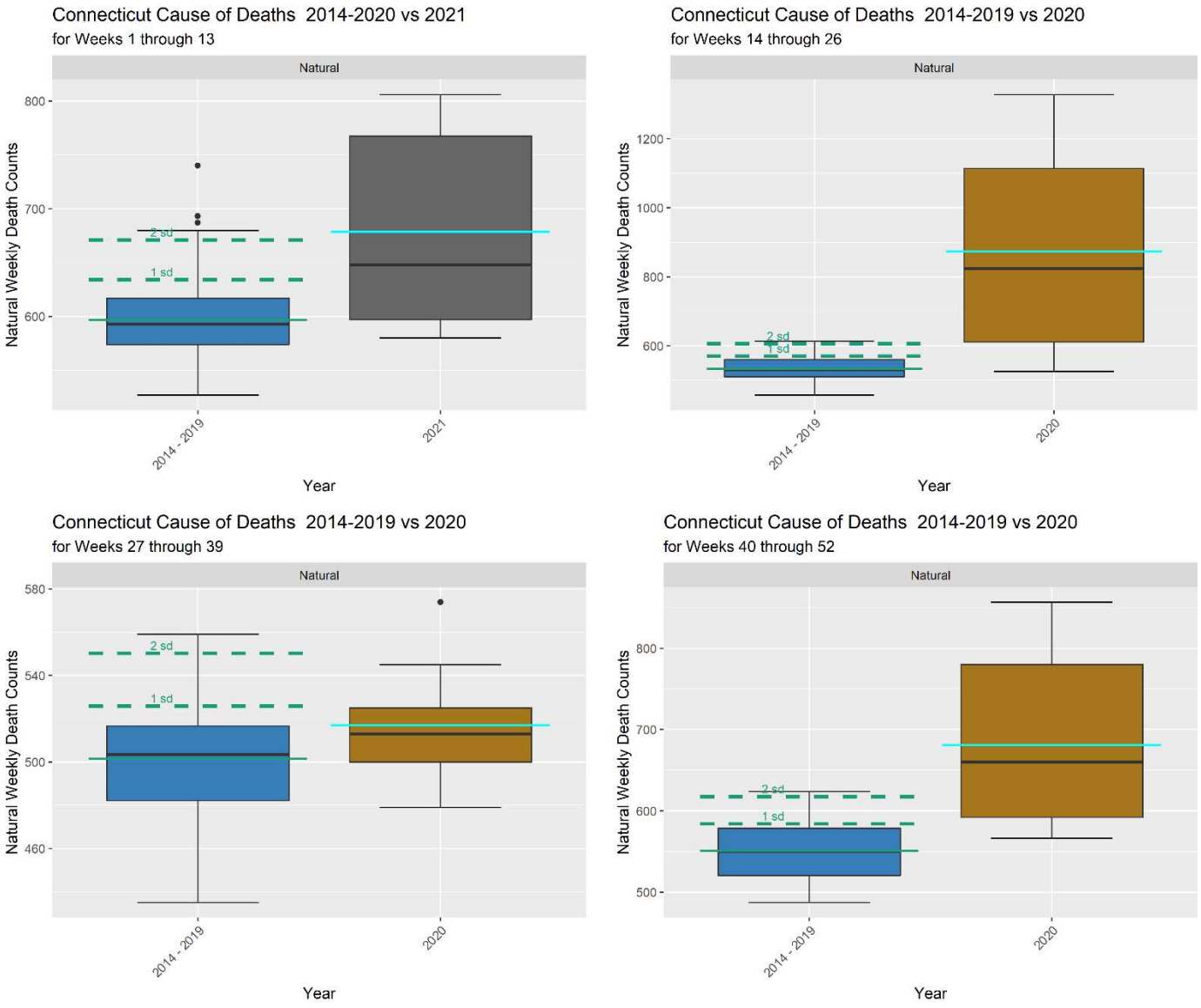


Based on the figure above, natural causes of death are increasing every three years showing a possible long-term trend. There a greater increase from 2019 to 2020 and several outliers exist in the 2020 data similar to other states' data. This is likely due to the same reasons as given previously, Covid-19 misdiagnosis or an aging population bubble. This further supports the idea that natural causes of death are a nationwide issue and may benefit from national level resources being applied to reduce the number of weekly deaths.





Both figures show the familiar pattern seen in the New Jersey and Massachusetts data. Natural cause weekly deaths increased significantly in the 2nd and 4th quarters of 2020 with the 4th quarter peak continuing into the 1st quarter of 2021. The figure on the right shows much lower growth and declines for the years 2015 through 2019 than in 2020 and 2021.



Based on the data shown in the figures above, the 2nd and 4th quarters of 2020 have an extreme significance level for natural causes of weekly deaths. The 1st quarter deaths of 2021 show a greater that 95% significance level but not as extreme as seen in the 2nd and 4th quarters.

Connecticut summary analysis and recommended course of actions

Based on the findings in the Connecticut natural causes of weekly death categories, I recommend that Connecticut healthcare systems follow the same recommendations given earlier and are listed below.

1. Work with their geriatric (elderly) populations and caregivers to ensure that their physical, nutritional, medical, emotional, and social needs are being met. This older population might be used to going it alone and may not reach out for help.
2. Focus long term resources on reducing natural causes of death and continue monitoring long-term trends.
3. Seek assistance and resources at the national level.
4. Collaborate with other states and regions to determine best practices and assess if they can be implemented in Connecticut. After implementing the changes, further monitoring of the data should be conducted to assess the effect on Connecticut's long-term weekly death trends.

Summary of Analysis, Recommendations and Conclusion

The 2020 and 2021 years' data shows a significant increase in weekly natural cause deaths when compared to previous years. The figures used in the Analysis and Recommendations section shows 95% or greater statistical significance for natural causes in all state data analyzed. This is possibly due to Covid-19 misdiagnosis or aging population bubble as with the Baby Boomers generation. Other causes of death disease appearing in the top 5 states with statistically significant weekly death growth rates are heart disease deaths, influenza, and pneumonia deaths.

Though not in the top 5 states data, diabetes was found to be prevalent in the top 20 states when natural cause of death was excluded.

Statistical Significance Methodology

To determine the states and regions with the most statistical significance, the data for all years previous to 2020 and 2021 is processed using dplyr's statistical summarize function with qnorm set to 68, 90, 95 and 99 percent significance level. This gives the means for the significance levels.

Using the 95th percentile means (2 standard deviations) as a quotient and the means of the 2020 and 2021 data as the numerator, a ratio is made. This ratio gives a factor or percent representing how many times greater than 2 standard deviations (95% significance level) the 2020 and 2021 mean value is. If the ratio is 0.5 then the mean is half of 2 standard deviations (1 std dev). If the factor is 1, the mean is at 2 standard deviations and if the factor is 3 it is at 6 standard deviations (3 times 2 std dev). This is done to make comparisons of different disease data easier by normalizing the data to 95%, the target significance level. It was later realized that a ratio of 1 std deviation would have made calculating the standard deviation slight simpler. Other functions and graphics were built based on the 2 standard deviation ratio. Therefore, the 2 standard deviation benchmark was not changed.

Success Criteria and Findings

The objective of the project is to analyze state level weekly cause of death data available from the Centers for Disease Control (CDC) to identify specific diseases with the highest growth in deaths so that it may be used to direct resources to reducing the number of deaths in the disease category.

The objective was met with a greater than 95% significance level. The states with the highest growth in weekly deaths are listed below including the standard deviations from the mean.

The Top 5 states or special regions showing the most statistically significant growth in disease weekly death rates based on a one-sided 95% significance level are:

1. New York City for natural cause weekly deaths in q2 of 2020 with the mean at 5.74 standard deviations from the previous years' mean. And heart disease weekly deaths for the 2nd quarter of 2020 with the mean at 3.58 standard deviations from the previous years' means.
2. New Jersey for natural cause weekly deaths in q2 of 2020 with the mean at 3.78 standard deviations from the previous years' mean. And influenza and pneumonia weekly deaths for the 2nd quarter of 2020 with the mean at 3.18 standard deviations from the previous years' means.
3. South Dakota for natural cause weekly deaths in q4 of 2020 with the mean at 3 standard deviations from the previous years' mean.
4. Massachusetts for natural cause weekly deaths in q2 of 2020 with the mean at 2.94 standard deviations from the previous years' mean.
5. Connecticut for natural cause weekly deaths in q2 of 2020 with the mean at 2.94 standard deviations from the previous years' mean.

Notable Findings

The objective of the project is to analyze state level weekly cause of death data available from the Centers for Disease Control (CDC) to identify specific diseases with the highest growth in deaths so that it may be used to direct resources to reducing the number of deaths in the disease category. The objective was met with a greater than 95% significance level.

The Top 5 states or special regions showing the most statistically significant growth in disease weekly death rates based on a one-sided 95% significance level are:

1. New York City for natural cause weekly deaths in q2 of 2020 with the mean at 5.74 standard deviations from the previous years' mean. And heart disease weekly deaths for the 2nd quarter of 2020 with the mean at 3.58 standard deviations from the previous years' means.
2. New Jersey for natural cause weekly deaths in q2 of 2020 with the mean at 3.78 standard deviations from the previous years' mean. And influenza and pneumonia weekly deaths for the 2nd quarter of 2020 with the mean at 3.18 standard deviations from the previous years' means.
3. South Dakota for natural cause weekly deaths in q4 of 2020 with the mean at 3 standard deviations from the previous years' mean.
4. Massachusetts for natural cause weekly deaths in q2 of 2020 with the mean at 2.94 standard deviations from the previous years' mean.
5. Connecticut for natural cause weekly deaths in q2 of 2020 with the mean at 2.94 standard deviations from the previous years' mean.

Recommended Course of Actions based on findings

For all states with statistically significant natural cause of weekly death growth, my recommendations are:

1. Work with their geriatric (elderly) populations and caregivers to ensure that their physical, nutritional, medical, emotional, and social needs are being met. This older population might be used to going it alone and may not reach out for help.
2. Focus long term resources on reducing natural causes of death and continue monitoring for flattening of the long-term trend.
3. Seek assistance and resources at the national level.
4. Collaborate with other states and regions to determine best practices and assess if they can be implemented in other regions or nationally. Analysis of the data should continue to validate positive effects on the long-term trend.

For New York's statistically significant heart disease weekly death growth, my recommendations are:

1. Continue to use and following current best practices already in place and monitor the data for any changes. The current practices are decreasing weekly deaths.

For New Jersey's statistically significant influenza and pneumonia weekly death growth, my recommendations are:

1. Focus short term resources towards reducing influenza and pneumonia deaths in the first quarters of the year. Where previous years' data showed a long-term pattern of significant increases.
2. Identify procedures and protocols that were effective during social distancing and promote the non-invasive ones that will be easily accepted by the public. A possible example is signage and reminders to wash hands frequently.

All states implementing these recommendations should monitor changes in their data sources on a quarterly basis to see if the results are having an effect on financial, material and labor capital costs.

Expected Benefits

By applying these recommendations, healthcare systems may reallocate resources to reduce waste in expenditures, materials, and labor. It will also have the effect of reducing death rates for diseases with the highest growth rate. The goal is to have a decrease in weekly death rates meeting a threshold of 90% or greater statistical significance level. Though, a smaller significance level such as 68% at 1 standard deviation could also be considered successful due to the lives saved and costs reduced.

Future Recommended Analysis and Studies

Recommendations for future study are:

- Continue the analysis with updated data to examine diabetes trends in the United States.
- Examine the relationship of seasonal viruses on some causes of death. There is a possible moderate correlation with Covid-19 and other disease such as brain and heart disease.
- Study seasonal patterns of non-seasonal diseases. The data shows seasonality for some deaths such as cancer and heart disease. Seasonal patterns seen in the data are weekly deaths rising during the fall and winter season, and decreasing during spring and summer months. Determining why these seasonal deaths occur could lead to treatments and strategies that reduce non-seasonal disease deaths.

Limitations in the data are the effect of missing data that was redacted due to privacy concern protections, delays in reporting of data and possible reporting and sampling biases. Another limitation is that states' weekly death rates for specific diseases should not be compared to each other due to participation rates in reporting varying by state. In addition, the data has cause of death reporting inconsistencies. These are seen in the natural death data that has a possible moderate correlation with Covid-19 possibly due to misdiagnosis or comorbidity. Causes of death where multiple causes are a factor could lead to a bias towards one diagnosis over another.

Sources

Centers for Medicare & Medicaid Services (2020, December 24). NHE fact sheet. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet.

Shrank, W. H., Rogstad, T. L., & Parekh, N. (2019). Waste in the US Health Care System: Estimated Costs and Potential for Savings. *JAMA*, *322*(15), 1501–1509. https://doi.org/10.1001/jama.2019.13978.

Sorato M.M., Asl A.A., & Davari M (2020). Improving Health Care System Efficiency for Equity, Quality and Access: Does the Healthcare Decision Making Involve the Concerns of Equity? Explanatory Review. *J Health Med Econ*., 6(1), 1-8. https://health-medical-economics.imedpub.com/improving-health-care-system-efficiency-for-equity-quality-and-access-does-the-healthcare-decision-making-involve-the-concerns-of.pdf.

Tikkanen, R., & Abrams, M. K. (2020, January 30). U.S. Health Care from a Global Perspective, 2019: Higher Spending, Worse Outcomes? U.S. Health Care from a Global Perspective, 2019 | Commonwealth Fund. https://www.commonwealthfund.org/publications/issue-briefs/2020/jan/us-health-care-global-perspective-2019.

U.S. Census Bureau. (n.d.) *International Data Base Population estimates and projections for 228 countries and areas.* Retrieved July 22, 2021, from https://www.census.gov/data-tools/demo/idb/#/table.

References

National Center for Health Statistics (2021). *Weekly Counts of Deaths by State and Select Causes, 2014-2019* (January 21, 2021) [Data set]. National Center for Health Statistics. https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-State-and-Select-Causes/3yf8-kanr

National Center for Health Statistics (2021). *Weekly Provisional Counts of Deaths by State and Select Causes, 2020-2021* (July 21, 2021) [Data set]. National Center for Health Statistics. https://data.cdc.gov/NCHS/Weekly-Provisional-Counts-of-Deaths-by-State-and-S/muzy-jte6

National Center for Health Statistics (2020, May 20). *Reporting and Coding Deaths Due to COVID-19*. National Center for Health Statistics. https://www.cdc.gov/nchs/covid19/coding-and-reporting.htm

Prabhakaran, S. (2019, November 13). *data.table in R – The Complete Beginners Guide*. Machine Learning Plus. https://www.machinelearningplus.com/data-manipulation/datatable-in-r-complete-guide/

Tutorial Point (n.d.). *R Tutorial*. Tutorial Point. https://www.tutorialspoint.com/r/index.htm

Seagull, J.L. (2019, August 22). *Order facet_wrap plots by descending order*. Stack Overflow. https://stackoverflow.com/questions/57613354/order-facet-wrap-plots-by-descending-order