An Investigation into the Correlation between Covid19-Related Deaths in 2020 and the Leading Causes of Death in the United States from 2015-2019

Authors: Maryam Aliakbari
Farzin Valiloo

Table of Contents

Table of Figures

List of Tables

1. Abstract

This report explores the death causes of different states in the U.S. for the years between 2014 and 2019 for the purpose of finding a relationship between top death causes in these states and the death rates of Covid-19 in 2020. In order to achieve this, we have utilized statistics models and machine-learning modeling. The statistics model shows a significant difference between the group of states which have different causes of death as their top death cause in terms of Covid-19 death ratios. The machine learning model that was chosen is Decision Tree model and after tuning this model, we've seen that the results are consistent with the statistics. The implication of these results can help states to be cautious when a possible future Covid break out may occur. The methods can also be optimized for other regions and countries if the proper data is available.

2. Introduction and Review

According to the World Health Organization (WHO), the top 10 causes of death accounted for 55% of the 55.4 million deaths worldwide in 2019 (World Health Organization, 2020).

The planning, development and sustainable implementation of health policies and health systems ought to be based on precise measurements and understandings of prevalence and incidence of communicable and non-communicable diseases, accidents, and other disabilities, given past and current demographic and epidemiological profiles in societies as well as how they are predicted to change over time (DEFO, 2014). In the start of the year 2020, Covid-19 caught the world off guard despite of prior warnings of some public health experts and commercial risk modelers (Clarke, 2022). It is concluded by the USA Center for Disease Control and Prevention (CDC) that underlying death causes can contribute to the mortality rate of Covid-19 (Center for Disease Control and Prevention (CDC) , n.d.).

In this article we aim to explore the select death causes in different states in the U.S. between the years 2014 and 2019 and find potential implications and/or relationships between these causes and the mortality of Covid-19 in 2020. In other words, in this report we are going to find an answer to this question:

> *"Is there a difference in mortality of Covid-19 among the states with different highest causes of death in prior years to 2020?"*

To achieve this, we will examine the data set that is provided by U.S. Government's open data website (U.S. Government, n.d.) and also utilizing complimentary datasets to optimize and/or normalize this data based on the aspect we will explore.

3. Methodology

In this report, the course of action to achieve the objective of the project (which is stated in the previous section) is as the followings:

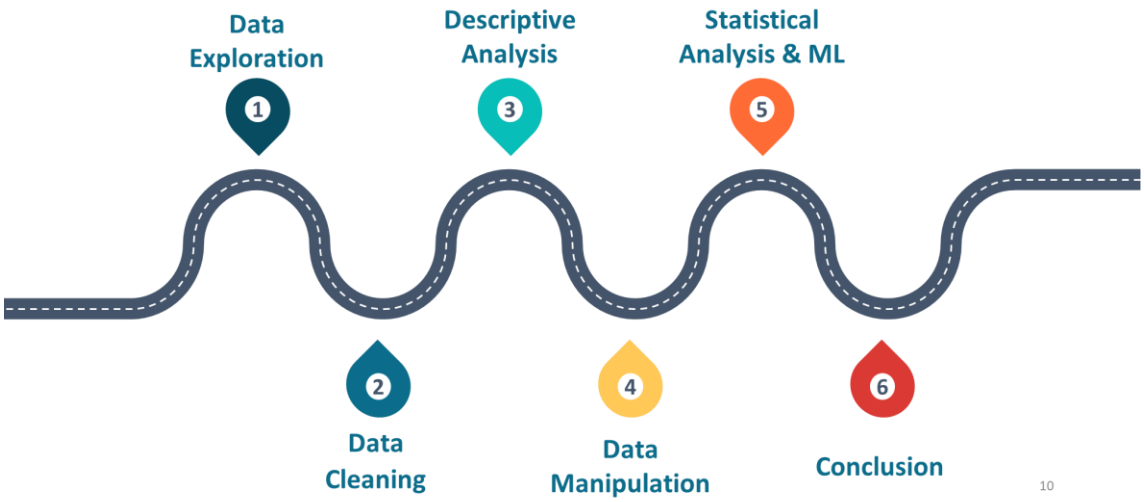


Figure 1- The overview of the roadmap in creating this report.

3.1. Data exploration

In this step we are going to explore the dataset to find any inconsistencies (missing data, excessive columns, wrongly formatted data , etc.). In order to do this, we are using MS Excel, Python and Tableau

3.2. Data cleaning

To address the potential issues from previous step, in this step we will use python to clean the dataset.

3.3. Descriptive analysis

By drawing few charts and plot using Python and Tableau, in this step we are going to get the first clear understanding of the data and how it is appropriate to answer the problem stated in the introduction section. The tools we'll be using for this step are Python and Tableau.

3.4. Data Manipulation

After we get a clear understanding of the data, we will understand what kinds of transitions and complimentary dataset we are going to need to attach to our dataset to be able to get to the answers we seek. The output of this step is a few new data sets that have just appropriate data to explore many aspects of the main question of the article. Python's Pandas and NumPy libraries are going to be the main tools utilized in this step.

3.5. Statistical analysis

We are going to utilize Python's SciPy library to statistically test hypotheses that we develop according to the articles question.

3.6. Application of a few machine-learning algorithms:

We are going to use a few machine learning algorithms to find out if any regression algorithms can get a proper score when being trained by the dataset. In this step, we are going to use PyCaret library to train and compare the regression models.

3.7. Conclusion

In this step we are going to summarize and conclude out findings and provide appropriate recommendations and/or insights into this subject.

- Tools: We will be using the following tools when going through this study:
  - Tableau 2022.3
  - MS Excel 365
  - MS Word 365 (to create this document)
  - Python 3.7.11
  - Pandas 1.3.5
  - NumPy 1.21.5
  - PyCaret 2.3.10
  - Matplotlib 3.5.3

4. Dataset

4.1. The main data set

As stated in the introduction section, the main data set is derived from U.S. government's open data website. This dataset contains 16,903 rows of data for the causes of deaths between 2014-2019. In each row we can see a week's number of deaths cause by on of the select death causes. In the following table you can see the columns of this data set:

US-LeadingCausesofDeath

Column name	Data type	Description
Jurisdiction of Occurrence	Text	State
MMWR Year	Numerical	The year part of the week
MMWR Week	Numerical	The week number in the year
Week Ending Date	Date	The date of the last day of the week
All Cause	Numerical	The overall number of deaths in the related week
Natural Cause	Numerical	The number of deaths caused by the related cause
Septicemia (A40-A41)	Numerical	The number of deaths caused by the related cause
Malignant neoplasms (C00-C97)	Numerical	The number of deaths caused by the related cause
Diabetes mellitus (E10-E14)	Numerical	The number of deaths caused by the related cause
Alzheimer disease (G30)	Numerical	The number of deaths caused by the related cause
Influenza and pneumonia (J10-J18)	Numerical	The number of deaths caused by the related cause
Chronic lower respiratory diseases (J40-J47)	Numerical	The number of deaths caused by the related cause
Other diseases of respiratory system (J00-J06,J30-J39,J67,J70-J98)	Numerical	The number of deaths caused by the related cause
Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27)	Numerical	The number of deaths caused by the related cause
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)	Numerical	The number of deaths caused by the related cause
Diseases of heart (I00-I09,I11,I13,I20-I51)	Numerical	The number of deaths caused by the related cause
Cerebrovascular diseases (I60-I69)	Numerical	The number of deaths caused by the related cause

Table 1- Table of columns in the main dataset

Please notice that the numbers in the “All causes” does not equal to the sum of other causes. The reason in probably because many deaths can be caused by a number of causes. Also there are a few more columns in the dataset that basically were dummy columns and did not have any analytical values so they were eliminated from the data set.

By exploring this data set, we faced the following issues:

Issue	Reason	Number of columns affected	Solution
Missing Values	When no deaths, the cell was left blank	11	Replaced by 0
Wrong data type	Comma (,) was misused	3	Deleted and replaced all commas and changed the value type

Table 2 - The issues encountered when exploring the main data set

4.2. Complimentary Data sets:

We have utilized a few complimentary datasets to have a deeper understanding of and address the problem statement. These datasets were clean and did not need any cleaning process in this stage.

4.2.1. US death causes 2020

To answer the problem, we also needed the number of deaths caused by Covid-19 in 2020. Fortunately, the same agency had a similar dataset for the causes of death for years between 2020 and 2022 (U.S. Government, n.d.) and the column for Covid-19 were add to the select death causes.

4.2.2. US Population 2010-2019

We also needed the population estimation of each state to normalize our data. Because census is held every 10 years. It was not possible to access the exact data, but we were able to find a reliable source for the estimation of population in each state for 2010-2019 (U.S. Census Bureau, n.d.) US Population 2020 and for 2020, (populationu.com, n.d.).

5. Descriptive Analysis

First, using the raw data (not normalized) the map below shows the distribution of death by all causes in the U.S. The map for 2014, 2019 and 2020 is showed to illustrate the changes that occurred by the spread of Covid-19.

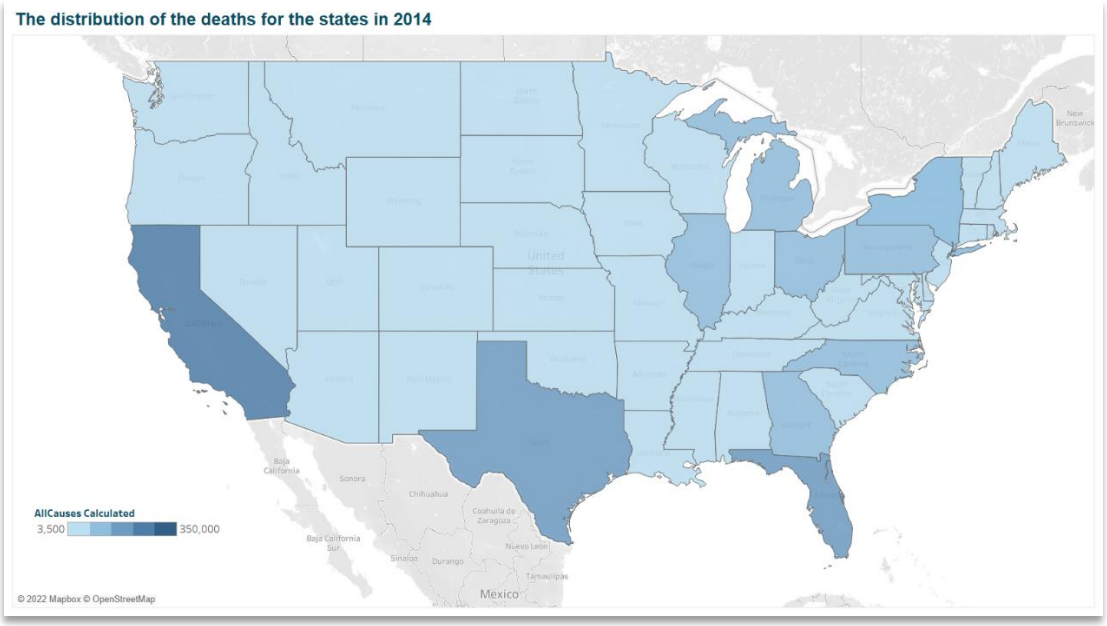


Figure 2- The distribution of the deaths for the states in 2014

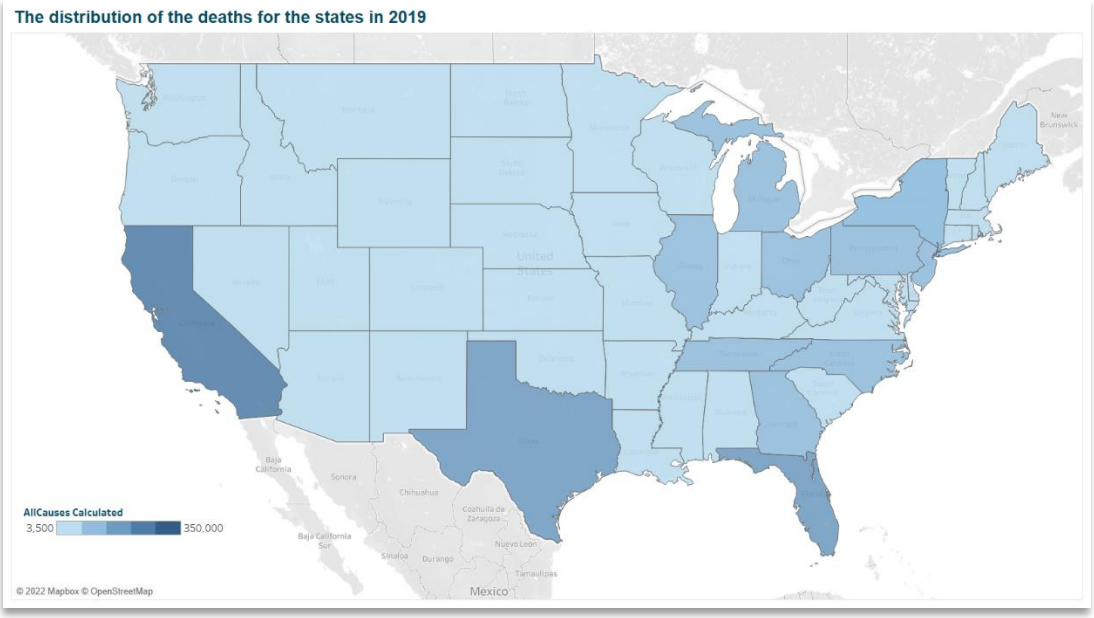


Figure 3- The distribution of the deaths for the states in 2019

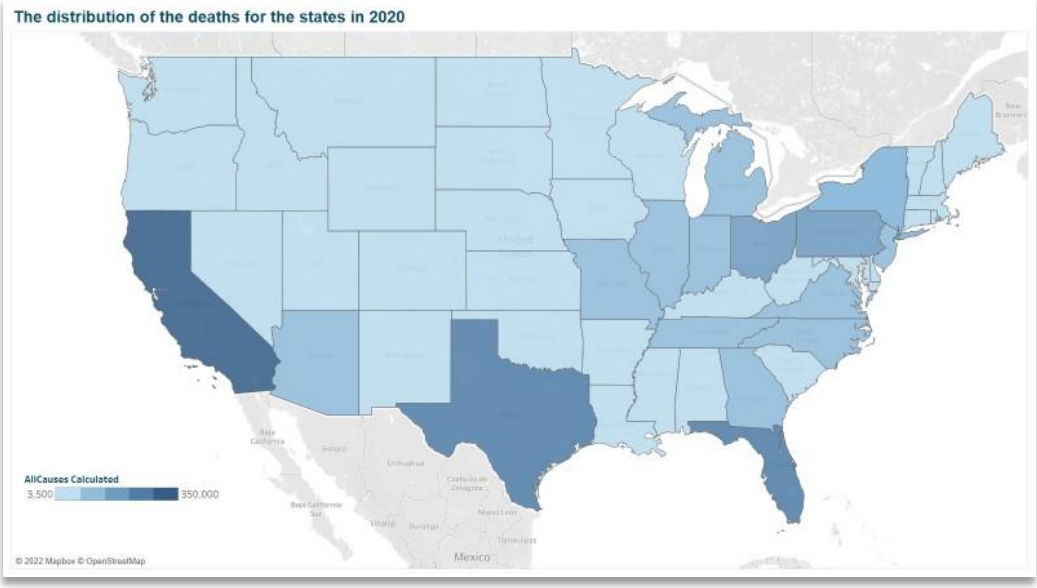
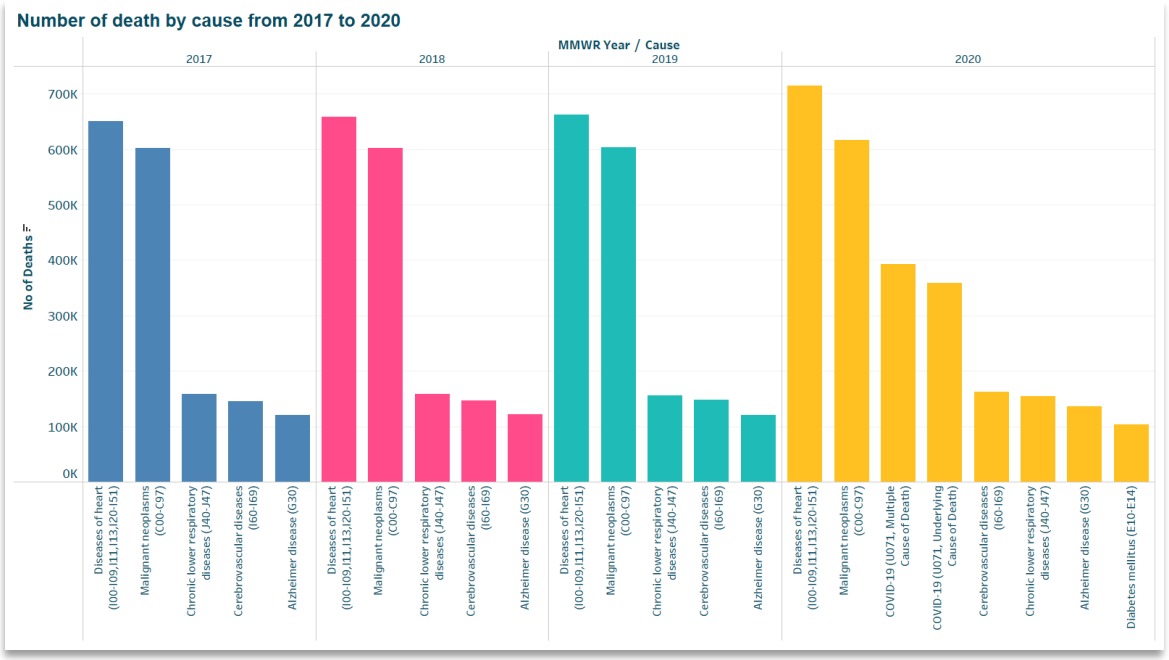


Figure 4- The distribution of the deaths for the states in 2020

Notice how colors become bolder when you compare the maps for 2019 and 2020.

The following charts also demonstrates the distribution of death causes in between 2014 and 2020.

US-LeadingCausesofDeath



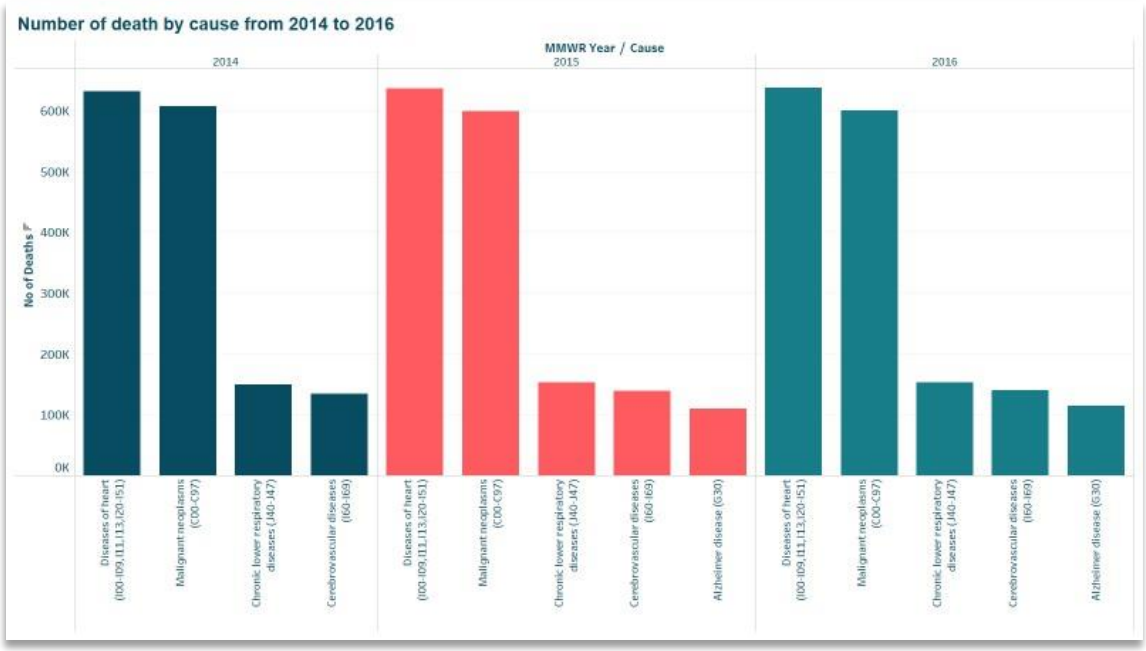


Figure 5- The causes that led to more than 100k deaths between 2014-2016 and 2017-2020

Notice that as the chart illustrates, in 2020 Covid related causes entered the view. For the years before 2020 the distribution of causes seems to be the same with minor differences.

6. Transformation of the dataset

In order to do a more precise analysis we need to normalize the data based on the population of each state. This way the illustrations will be clearer and can be understood and compared. To do this, we have accumulated the number of deaths for each cause in a yearly basis (because we only have the estimated population of states in a yearly fashion). For joining the datasets, the following issues were encountered:

Issue	Reason	Number of columns or rows affected	Solution
State names were different	Datasets were accessed through different data sources	All columns	Uniformed the names using MS Excel
One state missing in the population dataset	One dataset included all the territories of the U.S. not just the states	All columns and 7 rows	Eliminated correspondent data for the missing Territory

Table 3- The issues and respective solution for merging the datasets

After combining the datasets, we have divided the number of deaths for each cause by the population of the states in each year. Now let's look at the map of the U.S. based on the transformed data.

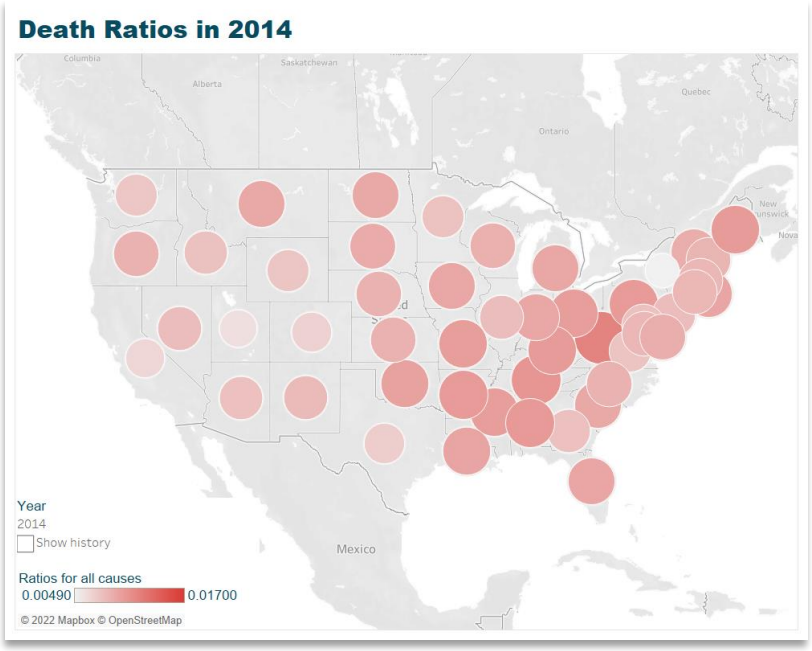


Figure 6- Map distribution of all death causes in the U.S. in 2014

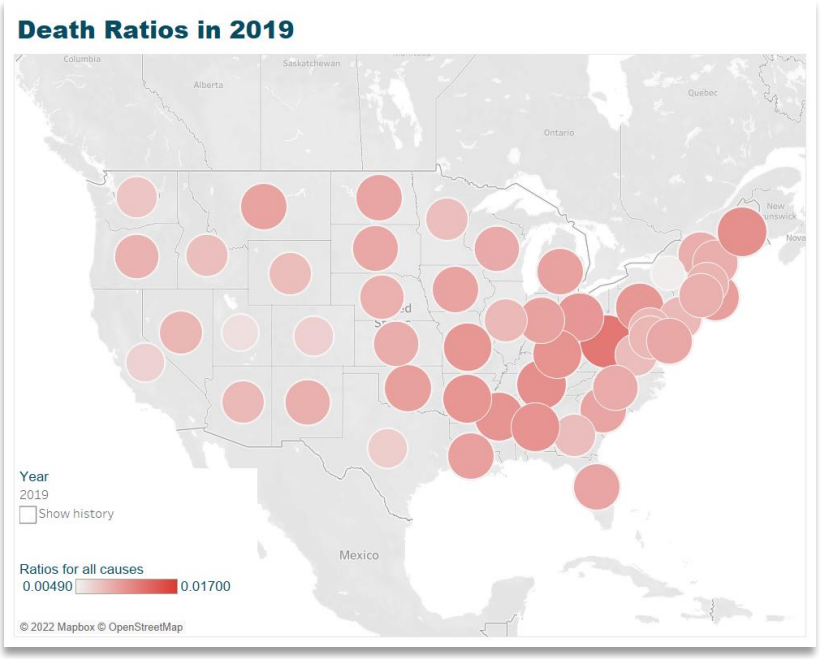


Figure 7- Map distribution of all death causes in the U.S. in 2019

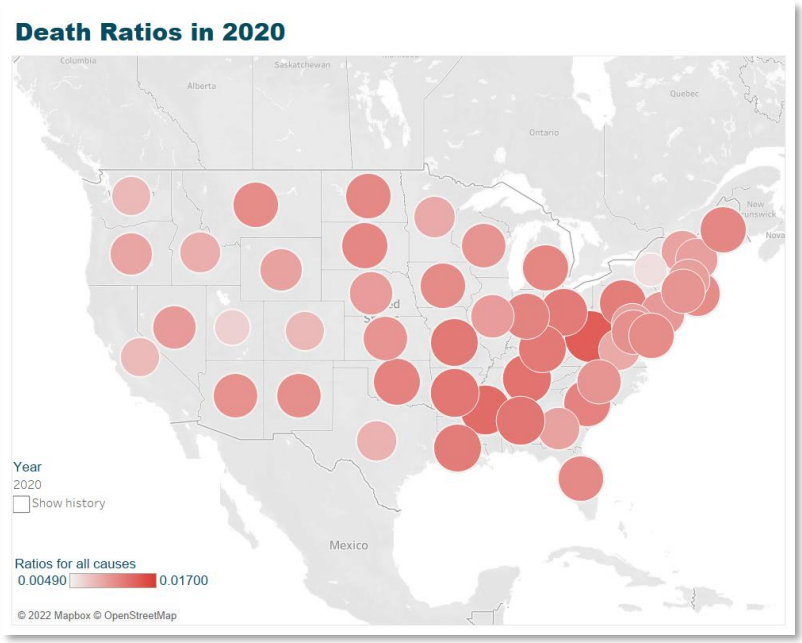


Figure 8- Map distribution of all death causes in the U.S. in 2020

As it can be seen in these charts, the death ratio between 2014 and 2019 is relatively similar (the boldness and size of the circles) but in 2020 these ratios have gone higher for almost every state.

US-LeadingCausesofDeath

This dataset is the base dataset for all the data analysis done in this project. All other transformations and data manipulations (related to the type of the analysis) is done based on this data set.

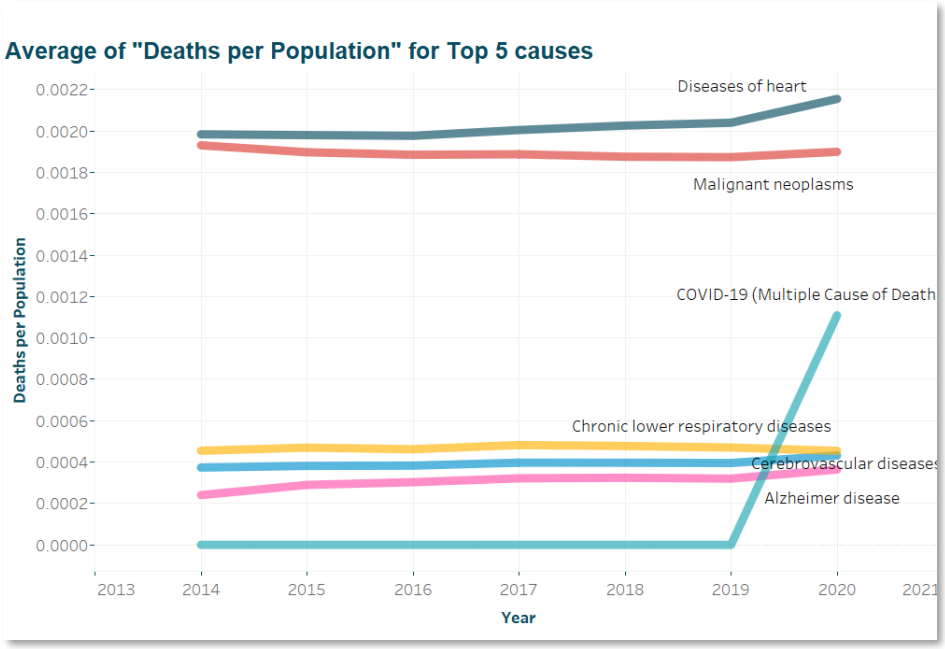


Figure 9- The trend of top death causes (per population) in the U.S 2014-2022

7. Analysis and results

"Do the states' Covid19 mortality rates (death ratios) affected by their top death causes in the previous years?". To answer this question (which is the main objective of this article) we need to:

- Accumulate the data for the years 2014-2019
- Create a list of the top death causes for each state in that period
- Compare the mortality rates for different states using various methods.

7.1. Statistical analysis

A data set has been created that contains top five causes of death for each state based on the average of normalized death rates in that state in the years 2014-2019. The two last columns in this data set contains the normalized death rate of Covid for the state. The data set looks like this:

US-LeadingCausesofDeath

	Jurisdiction of Occurrence	first_dc	second_dc	third_dc	fourth_dc	fifth_dc	COVID-19 (U071, Multiple Cause of Death)	COVID-19 (U071, Underlying Cause of Death)
0	Alabama	Diseases of heart (I00-I09,I11,I13,I20-I51)	Malignant neoplasms (C00-C97)	Chronic lower respiratory diseases (J40-J47)	Cerebrovascular diseases (I60-I69)	Alzheimer disease (G30)	0.001365	0.001275
1	Alaska	Malignant neoplasms (C00-C97)	Diseases of heart (I00-I09,I11,I13,I20-I51)	Chronic lower respiratory diseases (J40-J47)	Cerebrovascular diseases (I60-I69)	Septicemia (A40-A41)	0.000247	0.000195
2	Arizona	Diseases of heart (I00-I09,I11,I13,I20-I51)	Malignant neoplasms (C00-C97)	Chronic lower respiratory diseases (J40-J47)	Alzheimer disease (G30)	Cerebrovascular diseases (I60-I69)	0.001339	0.001232
3	Arkansas	Diseases of heart (I00-I09,I11,I13,I20-I51)	Malignant neoplasms (C00-C97)	Chronic lower respiratory diseases (J40-J47)	Cerebrovascular diseases (I60-I69)	Alzheimer disease (G30)	0.001360	0.001195
4	California	Diseases of heart (I00-I09,I11,I13,I20-I51)	Malignant neoplasms (C00-C97)	Cerebrovascular diseases (I60-I69)	Alzheimer disease (G30)	Chronic lower respiratory diseases (J40-J47)	0.000884	0.000827

Figure 10- Top 5 rows of the dataset created based on the top death causes in each state.

After analyzing the top death cause in every state, it is found out that two death causes (natural death cause is excluded from all the analysis) comprise the top two death causes in every state:

- Diseases of heart (I00-I09,I11,I13,I20-I51)
- Malignant neoplasms (C00-C97)

A comparison between the states whose first death cause was either is done visually:

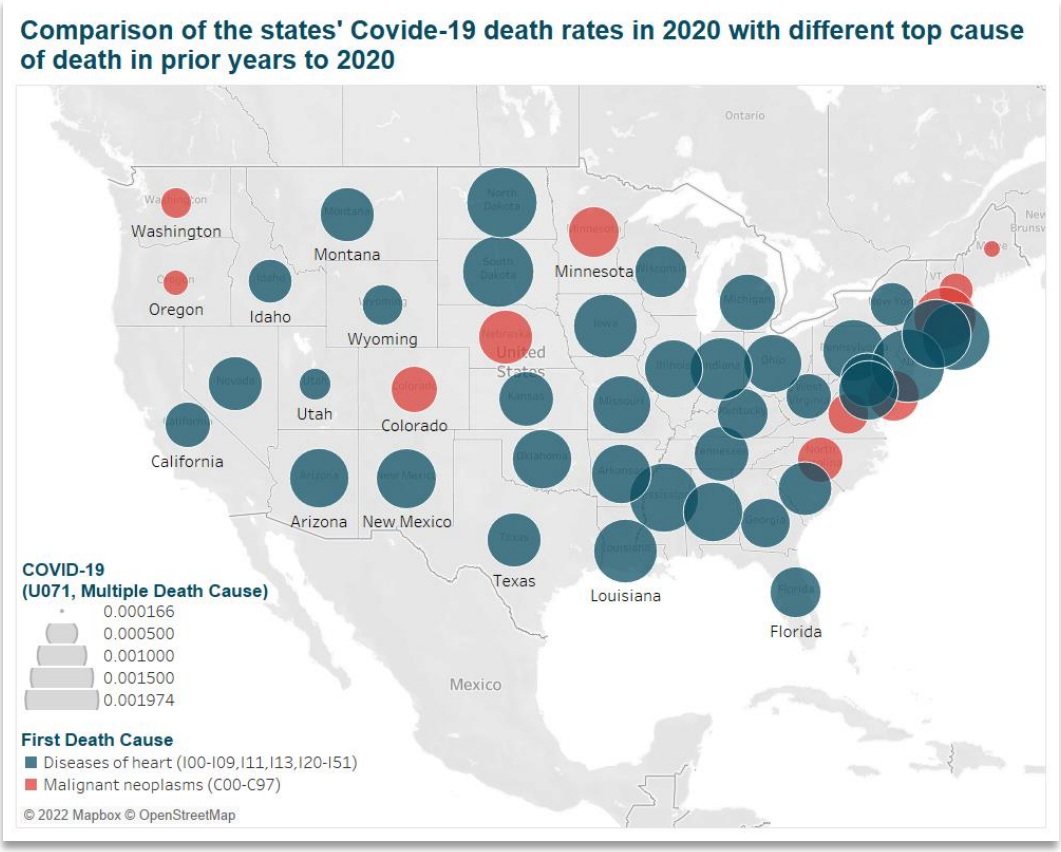


Figure 11- The death rates caused by Covid-19 in states with different top death cause prior to 2020

From this map it is likely that there might be a significant difference in the death rate caused by Covid-19 between these two groups. The chart below shows the distribution of this metric (death rate caused by Covid 19) in these two groups of states (states whose top death cause is “Diseases of heart” vs. the one whose top death cause is “Malignant neoplasms”):

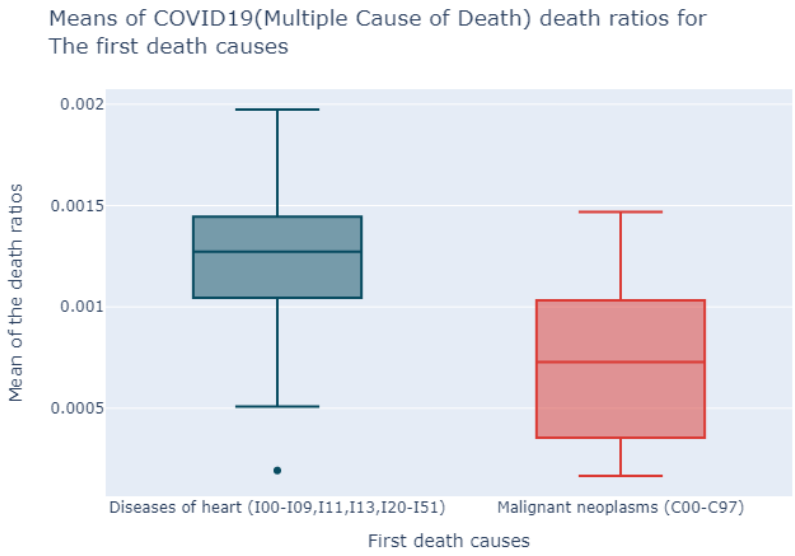


Figure 12- Comparison of the means of death ratios of COVID19(Multiple Cause of Death)

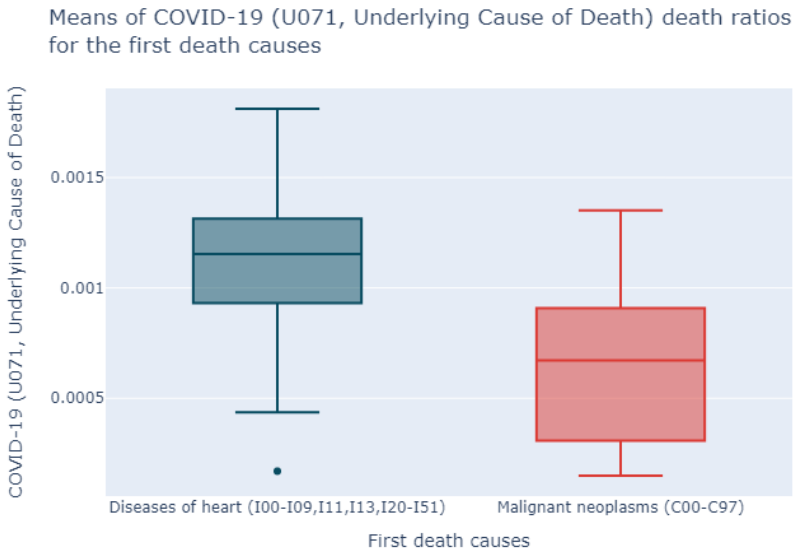


Figure 13- Comparison of the means of death ratios of COVID-19 (Underlying Cause of Death)

These two charts are consistent with findings in the previous figure. So, to test this claim we develop the following hypotheses:

Notice that we have two columns for Covid death rates: COVID19(Multiple Cause of Death) and COVID-19 (Underlying Cause of Death) and the exploration results are similar for both ratios. We are going to create a hypotheses test for both of these variables in the following part.

Group A states whose top death cause is “Diseases of heart”
Group B states whose top death cause is “Malignant neoplasms”

H0 *There is no difference in the means of death rates caused by COVID-19* ***(Multiple Cause of Death)***

H1 *The mean of death rates caused by COVID-19 (Multiple Cause of Death) in is higher in states group A than that of group B*

The result of t-test using SciPy library in Python:

Result

```
Ttest_indResult(statistic=array ([4.33382829]), pvalue=array([
3.62900604e-05]))
```

```
One-sided, Diseases of heart first:
 Ttest_indResult(statistic=array([4.33382829]), pvalue=array([3.62900604e-05]))
```

Figure 14- Python SciPy output for the first test (COVID19(Multiple Cause of Death))

P-value <0.01 So we can reject the null value in alpha = 99.99%

Now we are going to do this test for COVID-19 (Underlying Cause of Death):

Group A *States whose top death cause is “Diseases of heart”*
Group B *States whose top death cause is “Malignant neoplasms”*

H0 *There is no difference in the means of death rates caused by COVID-19* ***(U071, Underlying Cause of Death)***
H1 *The mean of death rates caused by COVID-19 (**U071, Underlying Cause of Death)** in is higher in states group A than that of group B than*

The result of t-test using SciPy library in Python:

Result

```
One-sided, Diseases of heart first:
Ttest_indResult(statistic=array([4.38565486]), pvalue=array([3
.0604979e-05]))
```

```
One-sided, Diseases of heart first:
 Ttest_indResult(statistic=array([4.38565486]), pvalue=array([3.0604979e-05]))
```

Figure 15- Python SciPy output for the second test (COVID-19 (Underlying Cause of Death))

P-value <0.01 So we can reject the null value in alpha = 99 %

Based on statistics analysis we can claim that Covid-19 caused higher death rate is the states that their top death cause in the previous years was “Diseases of heart”.

To continue to explore more we have investigated the second top cause of death for states are exact as the first top. In other words, the first and second cause of death for every state is the same (if the

rank is ignored). So, to go further we have decided to inspect the group A states and the third top cause of death in these states.

There are 38 states in group A and there are only two unique causes of death for these states (as the third top cause of death: “Chronic lower respiratory diseases (J40-J47)” and “Cerebrovascular diseases (I60-I69)”. The chart below illustrates the difference of mean of Covid-19 death ratio between two groups of states:

- The states with Chronic lower respiratory diseases (J40-J47) as their third top death cause (Group AC) and
- the states with Cerebrovascular diseases (I60-I69) as their third top cause of death (Group AD)

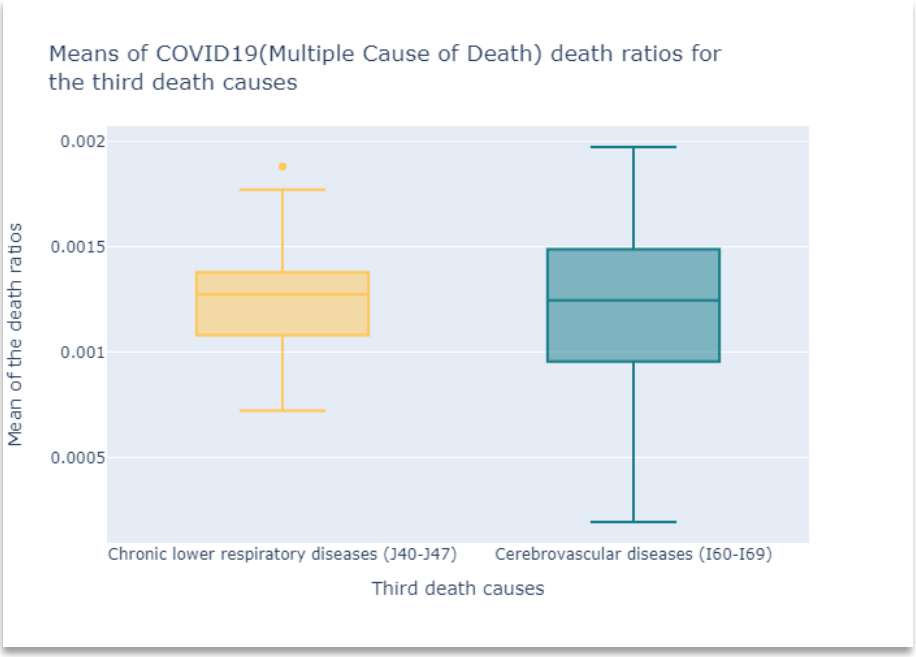


Figure 16- Figure 10- Comparison of the means of death ratios of COVID-19 (Multiple Cause of Death)

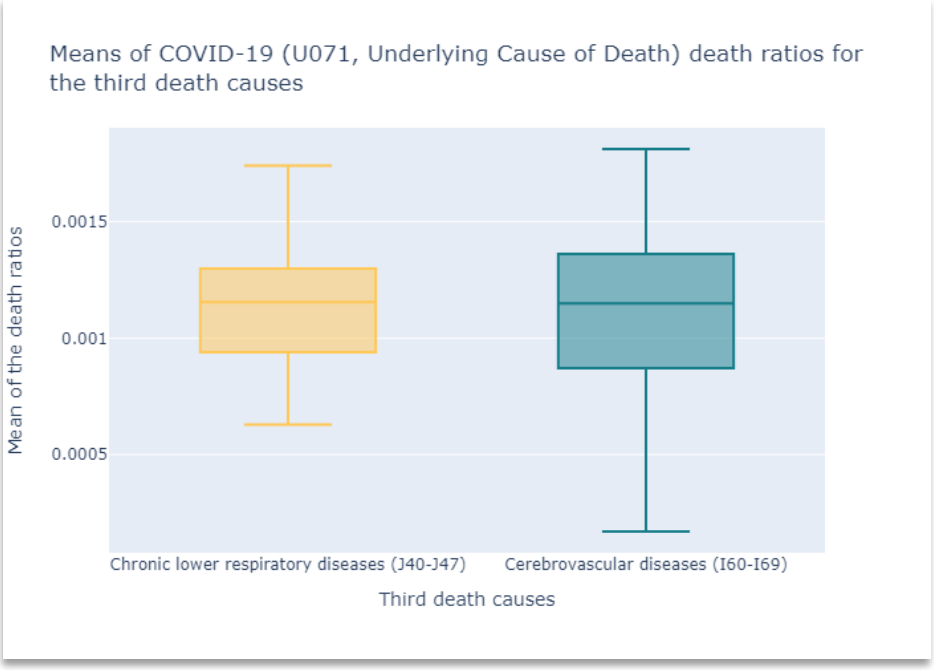


Figure 17- Comparison of the means of death ratios of COVID-19 (Underlying Cause of Death)

In the next step the t-test have been done. As before, two tests are going to be conducted for each of the dependent variable (COVID19(Multiple Cause of Death) and COVID19(Underlying Cause of Death)).

The hypotheses are:

Group AC | *States which their Third death cause is "Chronic lower respiratory diseases (J40-J47)"*
Group AD | *States which their Third death cause is "Cerebrovascular diseases (I60-I69)"*

H0 | *There is no difference in the means of death rates caused by COVID-19 (Multiple Cause of Death)*
H1 | *The mean of death rates caused by COVID-19 (Multiple Cause of Death) in is not equal in states group AC than and group AD*

The result of t-test using SciPy library in Python:

Result

```
Two-sided:

Ttest_indResult(statistic=array([-0.28025346]), pvalue=a
rray([0.78048737]))
```


```
Two-sided:
 Ttest_indResult(statistic=array([-0.28025346]), pvalue=array([0.78048737]))
```


Figure 18- Python SciPy output for the first test of third Cause (COVID-19 (Multiple Cause of Death))

P-value > 0.01 so we cannot reject the null hypotheses.

Now the second test for COVID-19(Underlying Cause of Death):

Group AC *States which their Third death cause is "Chronic lower respiratory diseases (J40-J47)"*
Group AD *States which their Third death cause is "Cerebrovascular diseases (I60-I69)"*

H0 *There is no difference in the means of death rates caused by COVID-19 (Underlying Cause of Death)*
H1 *The mean of death rates caused by COVID-19 (Underlying Cause of Death) in is not equal in states group AC than and group AD*

The result of t-test using SciPy library in Python:

Result

```
Two-sided:

 Ttest_indResult(statistic=array([-0.34385364]), pvalue=
array([0.73245812]))
```

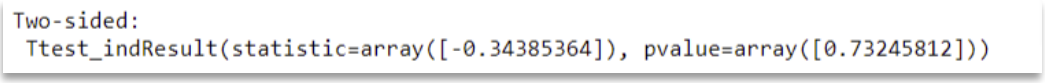


Figure 19- Python SciPy output for the for the second test of the third cause (COVID-19 (Underlying Cause of Death))

P-value > 0.01 so we cannot reject the null hypotheses

Based on these results there is no significant evidence that there is a difference in the average of deaths caused by Covid-19 between these two groups of states.

7.2. Utilizing Machine Algorithm

Thus, we were able to find out that: *"The average death ratio of Covid-19 in 2020 is different for the states with different top cause of death in prior years"*. In this section we aim to utilize a few machine-learning algorithm to explore potential relationship between previous top 5 death cause in different states and their death rates in 2020 (because of Covid-19). For this purpose, we will formulate these five top causes as feature for each state and Covid 19 death ratio as the target variable (because there is a high correlation between two variables that represent death rate of Covid-19 we will just use one of them as the target variable).

The tool we've chosen to implement machine-learning is PyCaret library for Python. The input dataset consists of the states as rows and the aggregation of the top five death causes as the features (the unique death causes that appeared as one of the top five death causes at least in one state). The vales are the average death rate of the cause in the respective states. The few top rows of the transformed dataset looks like as the following figure:

US-LeadingCausesofDeath

	d_of_heart	m_neoplasms	ch_low_resp	cerebvasc_d	alzheimer	septicemia	diabetes	influenza	COVID_19_2
0	2.645321	2.089085	0.681608	0.587634	0.486242	0.207753	0.242680	0.216043	1.275009
1	1.067281	1.232273	0.011735	0.004527	0.000000	0.000000	0.000000	0.000000	0.194985
2	1.725933	1.719721	0.532269	0.372256	0.422245	0.028486	0.294234	0.118915	1.231909
3	2.661376	2.158453	0.765488	0.510494	0.471708	0.083386	0.339679	0.175887	1.195408
4	1.567774	1.521659	0.343516	0.399841	0.395464	0.037583	0.234090	0.157392	0.826719

Figure 20- The overview of the dataset prepared as input for implementing M-learning models

In the next step we'll build and train regressor models in PyCaret. The results are as the following figure:

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
dt	Decision Tree Regressor	0.2616	0.1175	0.3161	-0.1487	0.1632	0.3816	0.0110
rf	Random Forest Regressor	0.2620	0.1240	0.3265	-0.1628	0.1620	0.4016	0.2690
knn	K Neighbors Regressor	0.3130	0.1638	0.3622	-0.2866	0.1835	0.5006	0.0390
omp	Orthogonal Matching Pursuit	0.2906	0.1570	0.3444	-0.3080	0.1714	0.4039	0.0120
br	Bayesian Ridge	0.2896	0.1681	0.3554	-0.4799	0.1807	0.4784	0.0090
et	Extra Trees Regressor	0.2834	0.1394	0.3521	-0.5277	0.1742	0.4068	0.2300
lasso	Lasso Regression	0.3299	0.1793	0.3924	-0.5417	0.1969	0.5230	0.0120
lightgbm	Light Gradient Boosting Machine	0.3299	0.1793	0.3924	-0.5417	0.1969	0.5230	0.0140
llar	Lasso Least Angle Regression	0.3299	0.1793	0.3924	-0.5417	0.1969	0.5230	0.0090
en	Elastic Net	0.3299	0.1793	0.3924	-0.5417	0.1969	0.5230	0.0110
dummy	Dummy Regressor	0.3299	0.1793	0.3924	-0.5417	0.1969	0.5230	0.0080
ada	AdaBoost Regressor	0.2664	0.1302	0.3378	-0.5525	0.1661	0.3940	0.0440
par	Passive Aggressive Regressor	0.2967	0.1500	0.3499	-0.5843	0.1860	0.3854	0.0120
gbr	Gradient Boosting Regressor	0.2711	0.1219	0.3361	-0.5981	0.1677	0.4054	0.0270
huber	Huber Regressor	0.2952	0.1788	0.3721	-0.6166	0.1870	0.4960	0.0130
ridge	Ridge Regression	0.3029	0.1761	0.3715	-0.8075	0.1872	0.4760	0.0100
lr	Linear Regression	0.3174	0.1930	0.3903	-1.0924	0.1955	0.4953	0.6570
lar	Least Angle Regression	0.3174	0.1930	0.3903	-1.0924	0.1955	0.4953	0.0100

Figure 21- PyCaret output after training different regressors model using the dataset.

As it can be seen in this figure, almost every model scores for this data set are relatively low. In order to investigate more and based on these scores we've chosen decision tree model to see if it can be tuned to create better results. The below figures show the tuned model of decision tree and its characteristics and results.

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.1743	0.0337	0.1837	0.2044	0.0997	0.2333
1	0.3395	0.2485	0.4985	-0.1689	0.3019	1.3330
2	0.2288	0.1018	0.3191	-0.4481	0.1483	0.1694
3	0.3291	0.1793	0.4234	0.0795	0.2368	0.6312
4	0.2807	0.1285	0.3585	0.3922	0.1647	0.2905
5	0.1025	0.0144	0.1200	0.3132	0.0531	0.0783
6	0.4009	0.2337	0.4834	-0.0312	0.2034	0.2873
7	0.1665	0.0393	0.1983	-0.1115	0.0961	0.1669
8	0.4618	0.2480	0.4980	0.5015	0.2603	0.9306
9	0.1952	0.0445	0.2109	-0.8227	0.1074	0.1782
Mean	0.2679	0.1272	0.3294	-0.0092	0.1672	0.4299
Std	0.1084	0.0895	0.1367	0.3834	0.0771	0.3885

Figure 22- The results for tuned Decision Tree Model after 10-fold validation

```
DecisionTreeRegressor(ccp_alpha=0.0, criterion='mae', max_depth=14,
                      max_features='log2', max_leaf_nodes=None,
                      min_impurity_decrease=0.02, min_impurity_split=None,
                      min_samples_leaf=4, min_samples_split=7,
                      min_weight_fraction_leaf=0.0, presort='deprecated',
                      random_state=1000, splitter='best')
```

Figure 23- The hyper parameters of the tuned Decision Tree model

As it is illustrated, the results have improved after tuning the model, but it is still far from optimal. The following figure illustrates the feature importance that has been extracted from this tuned model.

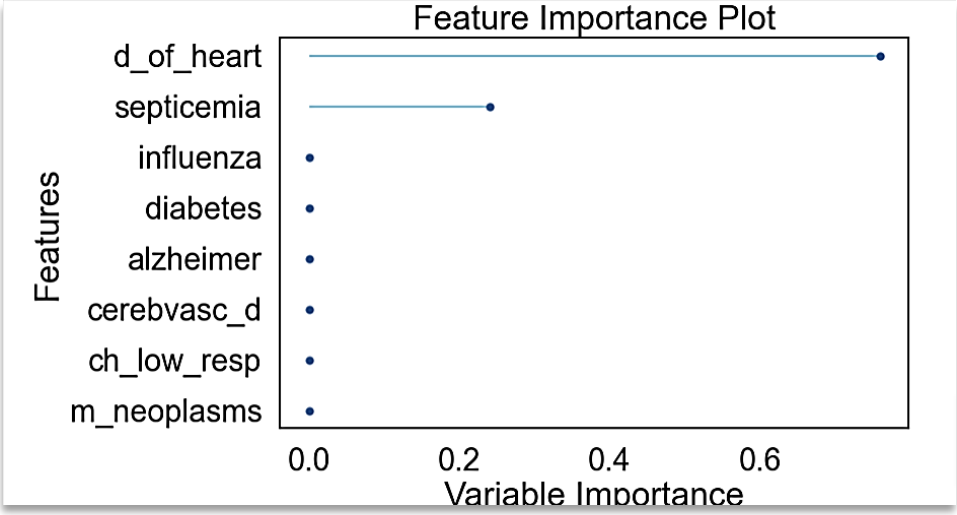


Figure 24- Feature importance extracted from the tuned Decision Tree Model

Although the scores of the decision tree model are not perfect, one thing that is noticed is that the most important feature extracted from tuning this model is the "Heart Diseases" which is consistent with the statistics model.

8. Discussion an implications

Although the results that were discovered in the previous section were consistent across the two methods that were utilized, but in the machine learning section, the model created had less than desirable scores. Some reasons that could have potentially affected these results are:

- The number of rows is too small for a machine-learning model to be able to train and be optimized
- The purpose of applying the model was to get a grasp of the feature importance in training data (which was achieved and was consistent with the other method), not to get an optimal score, because there was no data for test and the emergence of the Covid-19 was a onetime phenomenon.
- One of the reasons of this limited data was the nature of this study which needed many steps of data transformation and aggregation due to either the problem itself or the limited available data (specially periodic population data for each state). Considering these limitations findings are useful in few ways:
    1. Although vaccines and healthcare systems' adaptations have been relatively successful in dealing with Covid, but it is far from over (Harvard University, School of Public Health, n.d.). Thus, this study can be used to warn states in group A (whose leading

cause of death is “Diseases of Heart”) when cases of Covid or its variants have a possible surge in the future.

2. Among group A states, there were few states that had lower death rates because of Covid (e.g., Utah). The course of actions of the health care systems in these states can be investigated and learnt from by other states in this group.

9. Recommendations for future works

Researchers and analysts who may want to analyze this dataset can consider the following recommendations:

- Finding more detailed complimentary datasets that can be combined with this data will help to get a bigger picture of what can be done
- In future years the data about the Covid virus and its variants’ behavior will be available. This can be used to implement the logic in this study to find patterns in the relationship between death causes.
- The logic and methods used in this report can be used to compare the results with other countries and regions of the world (if the proper data is available) to conform or reject the findings in this report.

10. Contributions

Although this project was done by joint efforts of the two members of the group and no specific work was done entirely by a single member, but the following bullets summarizes the members' contributions:

Row	Tasks	Team Members	
		Maryam Aliakbari	Farzin Valiloo
1	Finding the dataset	●	●
2	Literature review	●	●
3	Data cleaning		●
4	Dataset transformation		●
5	Exploratory data analysis	●	
6	Quality control of the data sets and findings	●	●
7	Primary Visualization	●	
8	Statistics		●
9	Machine learning modeling	●	
10	Quality checking the codes and visuals	●	
11	Preparing final presentation	●	
12	Documentation and reporting	●	●
13	Creating final report		●

Table 4- Table of contributions of team members

11. References

- Center for Disease Control and Prevention (CDC) . (n.d.). *Excess Deaths Associated with COVID-19*. Retrieved from cdc.gov: https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm#dashboard
- Clarke, D. (2022, April 9). *Covid-19 caught the world off guard - pandemics must never surprise us again*. Retrieved from disasterprotection.org: https://www.disasterprotection.org/blogs/covid-19-caught-the-world-off-guard-pandemics-must-never-surprise-us-again
- DEFO, B. K. (2014). *Beyond the 'transition' frameworks: the cross-continuum of health, disease and mortality framework.* Retrieved from Global health action: https://doi.org/10.3402/gha.v7.24804
- Harvard University, School of Public Health. (n.d.). *The latest on the coronavirus*. Retrieved from hsph.harvard.edu: https://www.hsph.harvard.edu/news/hsph-in-the-news/the-latest-on-the-coronavirus/
- populationu.com. (n.d.). *US States by Population*. Retrieved from populationu.com: https://www.populationu.com/gen/us-states-by-population

- U.S. Census Bureau. (n.d.). *State Population Totals and Components of Change: 2010-2019*. Retrieved from census.gov: https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
- U.S. Government. (n.d.). *The home of the U.S. Government's open data*. Retrieved from Data.gov: https://catalog.data.gov/dataset/weekly-counts-of-deaths-by-state-and-select-causes-2014-2018
- World Health Organization. (2020, December 09). *The top 10 causes of death*. Retrieved from www.who.int: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

12. Appendices

All appendices listed below can be found in this link:

- https://github.com/maryamaa/US-LeadingCausesOfDeath

Row	File Name	File Format
1	DAB304-S3-Group1-FinalProject _Part1	.ipynb
2	DAB304-S3-Group1-FinalProject _Part2	.ipynb
3	DAB304-S3-Group1-FinalProject-DeathCausesDistibution	.twbx
4	DAB304-S3-Group1-FinalProject-Top5DeathCauses	.twbx
5	DAB304-S3-Group1-FinalProject-Maps	.twbx
6	DAB304-S3-Group1-FinalProject-1stTo3rdDeathCauses	.twbx
7	Weekly_Counts_of_Deaths_by_State_and_Select_Causes__2014-2022	.csv
8	PopulationUS	.csv
9	death_causes_population	.csv
10	death_causes_population_ratios	.csv
11	covid_start	.csv
12	covid_numerical	.csv

Table 5- Table of appendices files