The mortality rate of different age groups in the states of USA due to COVID19 since May to July have changed?

Neeharika Sinha

August 1st 2020

Table of Contents

1. Introduction

Tragically people have lost their lives due to COVID-19 all over the world. This pandemic started in the beginning of 2020 for United States. The infection being air borne, it was suggested to avoid human gatherings to prevent it spread. The death due to COVID-19 began reporting since April.

After various preliminary exploratory data analysis, the initial observation in February 2020 indicated that older people, and people with pre-existing medical conditions (such as asthma, diabetes, heart disease) appear to be more vulnerable to becoming severely ill with the virus. But with time it has been observed that almost all age groups got affected by the new coronavirus (2019-nCoV). The disease caused by the novel coronavirus has killed at least 159,000 people in the United States since February.

Being a data scientist, I wanted to understand how these deaths are distributed across age categories in the US population with time.

The Interpretation is to consider our grandparents, adults and kids keep up positivity and stay safe at home. But the real question is how we protect those that we love and get safely get back to work. This might be a small step to answer the question of how to protect lives and get the economy back up and running by looking at this data of the United States.

2. Data Acquisition and Wrangling

2.1 Source of dataset

The Centers for Disease Control and Prevention (https://www.cdc.gov/) website enumerates the data set for deaths involving coronavirus disease 2019 (COVID-19), pneumonia, and influenza reported to NCHS by sex and age group and state and is updated every week. The data was collected from the cdc website (https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku/data). Number of deaths reported in this table were the total number of deaths received as of the end of the week as the date reported.

This project includes data from the week of the month May, June, and July. Below are the dates when the cdc.gov website got updated.

5/6/2020,5/13/2020,6/3/2020, 6/10/2020,6/17/2020,6/24/2020, 7/1/2020', 7/8/2020, 7/22/2020, 7/29/2020

The xlsx format was preferred for the download and converted to data frame. Figure 1. shows the information of the data frame.

```
In [83]: df1.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1416 entries, 0 to 1415
Data columns (total 13 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Data as of                                  1416 non-null   datetime64[ns]
 1   Start week                                  1416 non-null   datetime64[ns]
 2   End Week                                    1416 non-null   datetime64[ns]
 3   State                                       1416 non-null   object
 4   Sex                                         1416 non-null   object
 5   Age group                                   1416 non-null   object
 6   COVID-19 Deaths                             1108 non-null   float64
 7   Total Deaths                                1223 non-null   float64
 8   Pneumonia Deaths                            1095 non-null   float64
 9   Pneumonia and COVID-19 Deaths               1082 non-null   float64
 10  Influenza Deaths                            888 non-null    float64
 11  Pneumonia, Influenza, or COVID-19 Deaths    1068 non-null   float64
 12  Footnote                                    927 non-null    object
dtypes: datetime64[ns](3), float64(6), object(4)
memory usage: 143.9+ KB
```

Figure 1 the information of the data sheet downloaded on weekly basis.

Among the various features provided, I was interested in the columns " Data as of the date", "Age_group" and "COVID-19 Deaths. The aim is to get to know if the average death due to COVID0-19 has differed through the months of May, June and July for 52 Staes and Union Territories of the United States. The xlsx sheet also consists of the overall data for the US as shown in Figure 2.

In [84]: df1.head()

Out[84]:

	Data as of	Start week	End Week	State	Sex	Age group	COVID-19 Deaths	Total Deaths	Pneumonia Deaths	Pneumonia and COVID-19 Deaths	Influenza Deaths	Pneumonia, Influenza, or COVID-19 Deaths	Footnote
0	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	Under 1 year	4.0	3951.0	36.0	1.0	11.0	50.0	NaN
1	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	1-4 years	2.0	780.0	33.0	2.0	33.0	66.0	NaN
2	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	5-14 years	4.0	1146.0	38.0	0.0	41.0	83.0	NaN
3	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	15-24 years	48.0	6843.0	143.0	18.0	41.0	211.0	NaN
4	2020-05-06	2020-02-01	2020-05-02	United States	All Sexes	25-34 years	317.0	14629.0	496.0	134.0	133.0	800.0	NaN

Figure 2. The features and columns provided in the data set

There was series of cleaning and data wrangling process, which included identifying the missing values, pulling only the overall deaths and not related to sex as can be seen from this IPython notebook. We had 11 data frames with formatted columns and rows. The nan indicates either no value was observed or the death was not reported. The nan values were replaced by '0' indicating no deaths.

3. Data Exploration

Since the beginning of the pandemic it has been predicted and proved many time that only elderly (>65 years) and those with underlying health conditions are more susceptible to the COVID infection.During January 1, 2020–May 18, 2020, approximately 1.3 million cases of coronavirus disease 2019 (COVID-19) and 83,000 COVID-19–associated deaths were reported in the United States [1].

By late June, however, the virus was coursing through a younger population. On June 20, Gov. Ron DeSantis (R) said the median age of a Floridian with covid-19 was down to 37. Although there has been reports or exceptional cases for younger generation getting the infection, luckily there has been reports for its recovery too. This might be due to strong immunity for the younger age group.

The COVID-19 deaths average age for all 50 states for each week (total 11 week) through May to July was calculated as shown in this IPython notebook.

The average age of COVID death for US for every 11 week was calculated from the data frame. Each of the range of age group was replaced by a mid value to get the average age for each state for each week as shown in Table 1. The Figure 1 plots of average death for mid value of the the age group ranging from 0 to >85 years for the week on May 6th 2020. The red line shows the average age for US for that week .

Age_group	Mid value
less_than_1_year	0.5
1_to_4 years	2.5
5_to_14 years	9.5
15_to_24 years	19.5
25_to_34 years	29.5
35_to_44 years	39.5
45_to_54 years	49.5
55_to_64 years	59.5
65_to_74 years	69.5
75_to_84 years	79.5
more_than_85 years	92.5

Table1: The age group reported in the data was replaced by a mid value

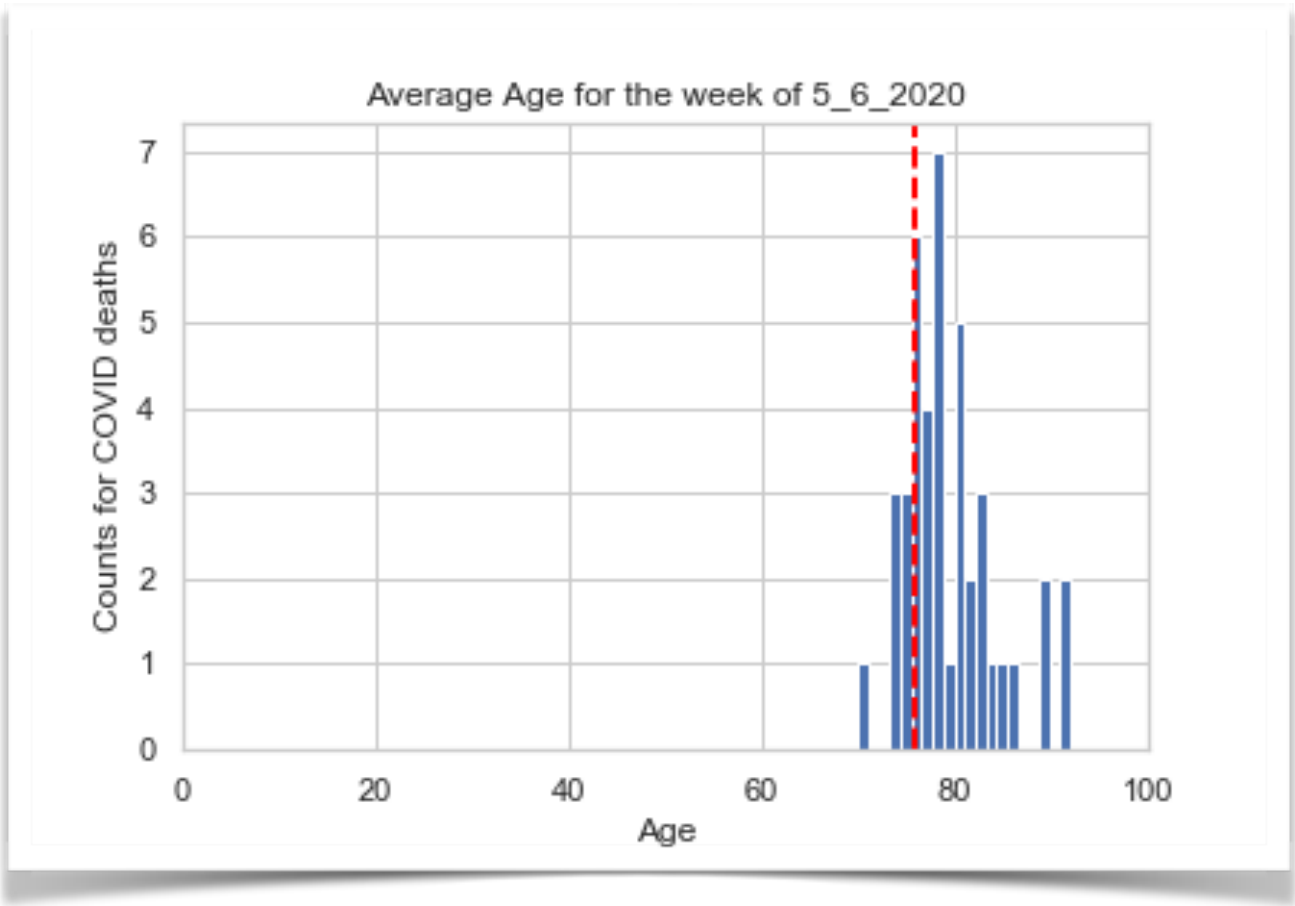

Figure 1. Average death for age group ranging from 0 to >85 years for the week on June 5th 2020

Similar calculations was carried for all the 11 weeks starting from the month of May , June to July as in this IPython notebook. The Figure 2 plots the average death for age group ranging from 0 to >85 years for 11 weeks starting from June 5th 2020 to July 29 2020. The red line shows the average age for US for each week .

As can be seen in the plot, the magnitude of COVID-19 deaths for the month of May is less as compared to June and July. Which very well explains that we in US were in the rise curve of the pandemic and the cases of deaths we picking up towards the beginning of June .

The histogram for every week as a somewhat uniform distribution with peak at around 77 years but skewed towards higher age group (>70 years). We can see few outliers for the week of 7th and 29th July. The analysis shows that these data points belongs to Puerto Rico, where we had reports of death in the age group of 50 years. The red line again marks the US average age group most effected my the virus.

In Figure 2 the plots were arrange so that each row corresponds to a month. The last block is kept deliberately empty. There is no missing data during the entire time of data collection. It was only that the cdc.gov website got update on June 3rd after May 28th 2020.

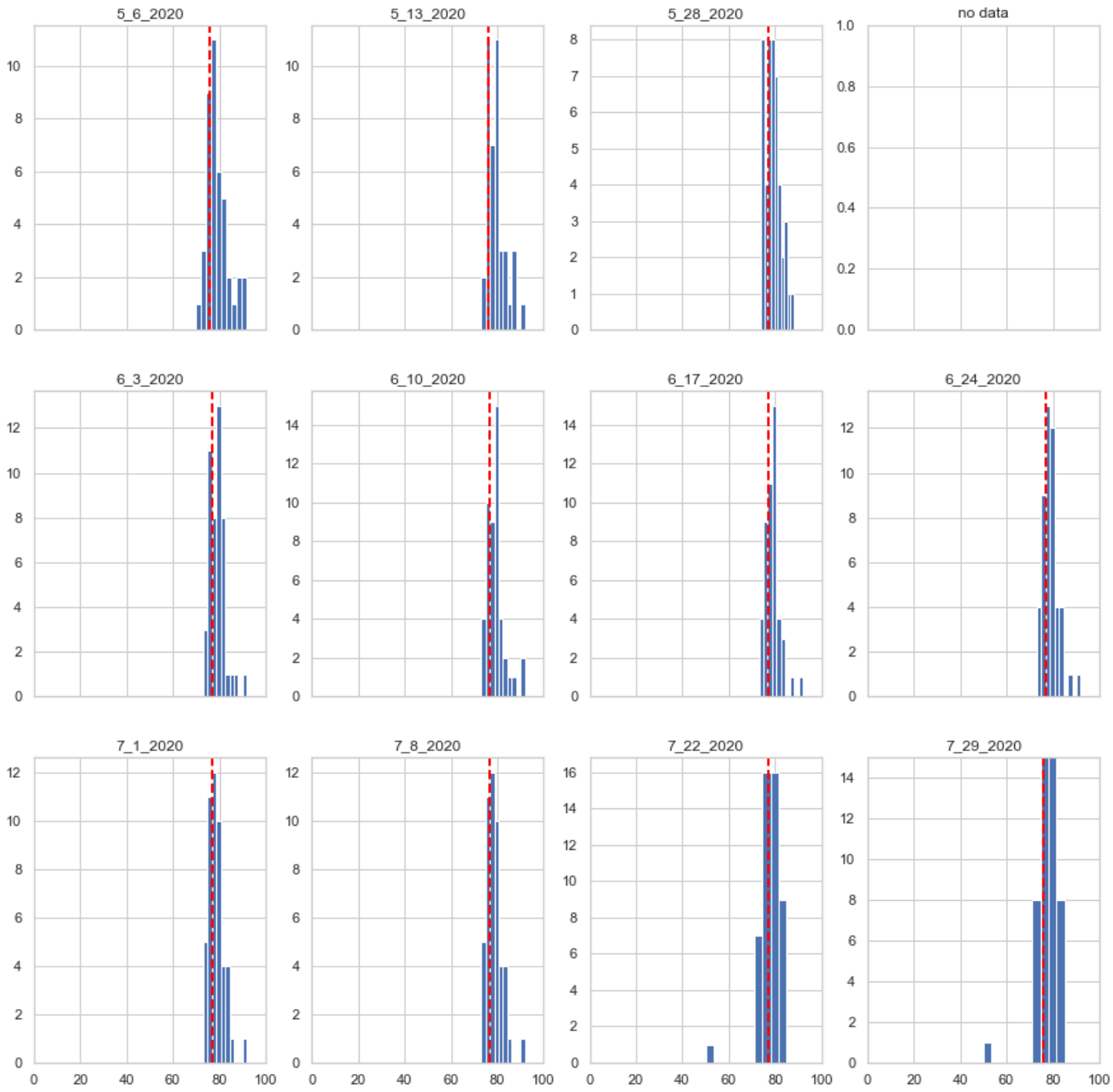

Figure 2 plots of average death for age group ranging from 0 to >85 years for 11 weeks starting from June 5th 2020 to July 29 2020. The red line shows the average age for US for each week

4. T statistic

4.1 Data Pre-processing

The average age group for COVID deaths through time was analyzed with every 11 weeks since May to July. Considering the hypothesis testing as

H0 = the average death is same through two weeks taking first week as 6th May 2020 and second week as all the other 10 weeks (i.e 5/13/2020,5/28/2020,6/3/2020,6/10/2020, 6/17/2020, 6/24/2020, 7/1/2020, 7/8/2020, 7/22/2020, 7/29/2020).

Ha = the average death have changed through two weeks

As the sample was selected from different states for different week from the same US population, the two sample independent t test was carried out for the combinations of the weeks. For the starting week of 5/6/2020 and the weeks of 5/28/2020, 6/3/2020, 6/17/2020, 6/24/2020, 7/1/2020, 7/8/2020, 7/22/2020 the p value is more than 5% and hence the null hypothesis is accepted and the alternate hypothesis is rejected as seen in this IPython notebook. So the average age for mortality through the week of 5/6/2020 to 7/22/2020 remained the same.

The t test failed for the immediate next week of 5/6/2020 that is 5/13/2020, as the data was not enough for the testing at the beginning of the pandemic in the month of May.

This shows that contrary to earlier consideration that only elder people are more prone to the COVID 19 virus, the virus eventually spread high enough to effect even the younger generation [2], but the average age for the death remains the same as the t test carried out in this project.

For the week of 7_29_2020 the null hypothesis can be rejected as the p value is less than 5%, so we can say that in recent months of July and August the average age for the COVID-19 deaths have changed. The outliers present in the week of 7_29_2020 is the specific reason for it.

Although the young generation got effected and there were reports of mortality, the average average group, the variation in the average age through the months of May, June and July can be tested by measuring the confidence interval as done in this IPython notebook.

Taking the difference of the average age group for all states and union territories, for consecutive weeks the variation in the average age through all states was analyzed. With a 95% confidence interval we can say that the significance of a shift of the average age for US over time, in this 11 weeks lies between 0.7 years to 10 years.

One point to be noted is the confidence interval was calculated including and removing Puerto Rico which can be considered as an outlier in our data set. The IPython notebook gives this analysis and while including Puerto Rico the variation had a difference on 1 year.

5. State Analysis

The total deaths for three separate months , May June and July was analyzed for 50 states namely (Alabama, Alaska , Arizona , Arkansas , California ,Colorado,Connecticut , Delaware, Florida , Georgia , Hawaii, Idaho , Illinois , Indiana , Iowa , Kansas , Kentucky,Louisiana, Maine , Maryland , Massachusetts , Michigan, Minnesota,Mississippi,Missouri, Montana , Nebraska , Nevada , New Hampshire , New Jersey , New Mexico , New York, North Carolina , North Dakota , Ohio , Oklahoma , Oregon , Pennsylvania , Rhode Island , South Carolina , South Dakota , Tennessee , Texas , Utah , Vermont , Virginia , Washington , West Virginia , Wisconsin , Wyoming).

5.1 Data Pre-processing

The average age of mortality for the end week of May, June and July plotted on the US map to have a view of the state variation. Again the mid value for the given age group was considered for the calculation as in Table 1 as in this IPython notebook.

The data reported in the cdc.gov website is a cumulative data for each week. So the idea of taking the last week of every month can be considered as the total death in that month.

5.2 Data Analysis and Visualization

The Figure 3 gives the average age distribution for each state with mortality reported for the month of May due to COVID-19. The plot shows that the majority of the states had the age group above 70 years. This average value is found to be very consistent even for the month of July as has been proved above with the t test. Figure 4 and 5 gives the distribution of average age group for each state with deaths reported for the month of June and July respectively.

The states Idaho, Arkansas, Minnesota, West Virginia and New Hampshire, the average age for Covid-19 was above 82 years age group. In the month of June the average age reduced and was reported only for the Idaho and North Dakota. Idaho has been reporting loss of old age group throughput the three months. It might be the reason that the old age population may be dominating.

The month of June has seen a wide spread of age group but for July majority of the states have a uniform average age for Covid -19 mortality.

Although it was reported that the infection is more susceptible to old age but through days the infections spread through different age group. We can conclude that although the infection spread for different age , the death rate still remained consistent through the months of May, June and July of around 77 years +- 10 years with 95% confidence.

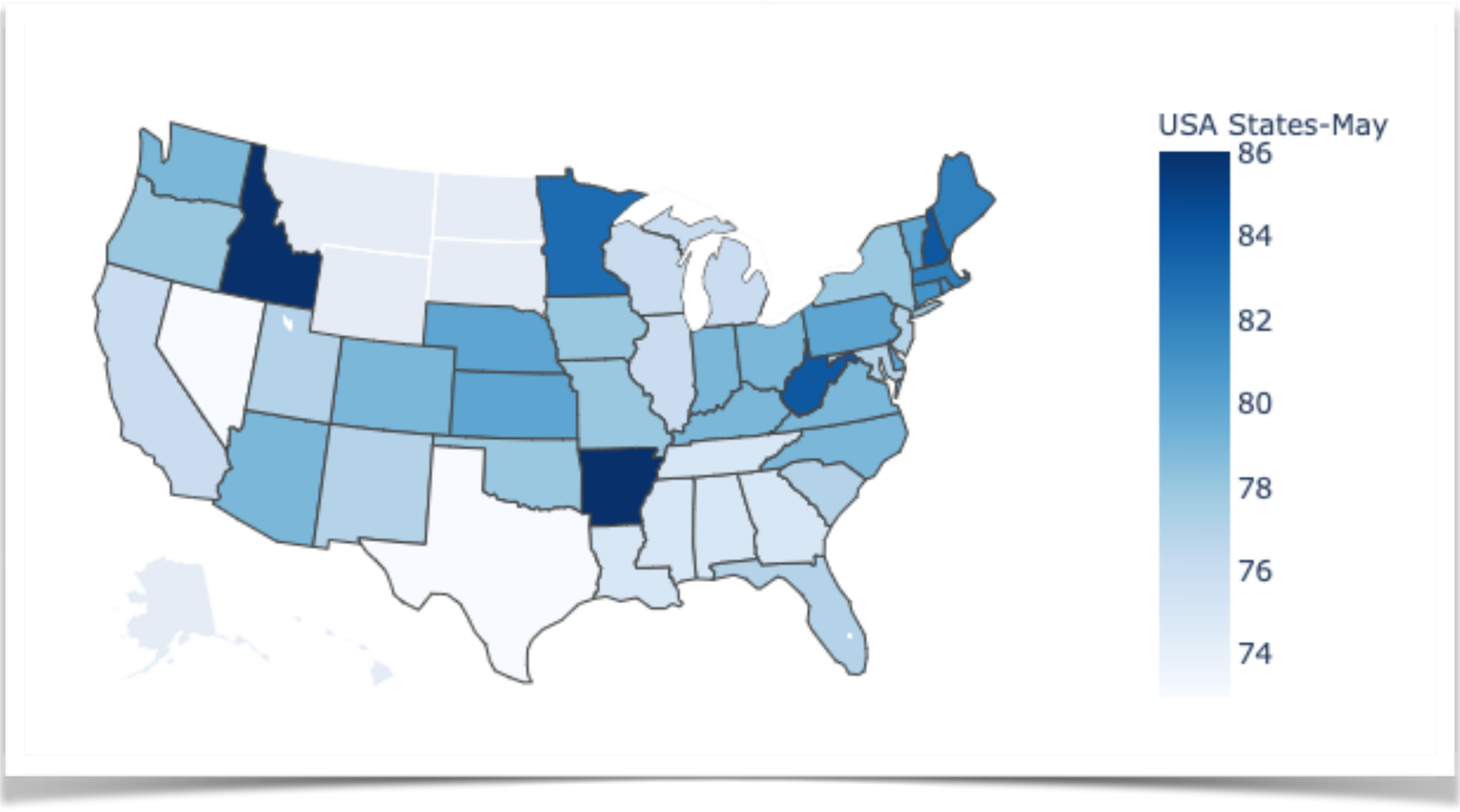

Figure 3: The distribution of average age group for each state w for the month of May

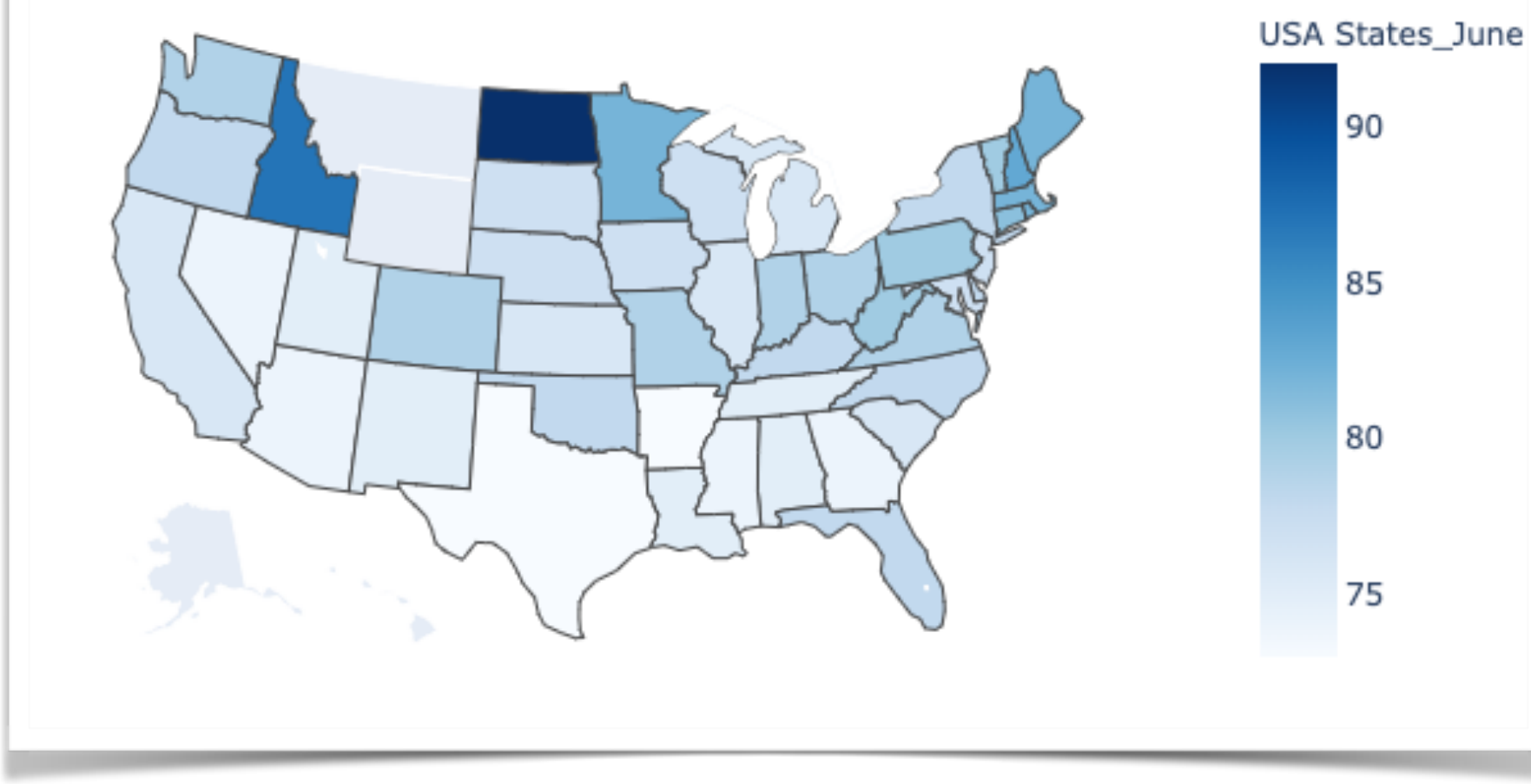

Figure 4: The distribution of average age group for each state w for the month of June

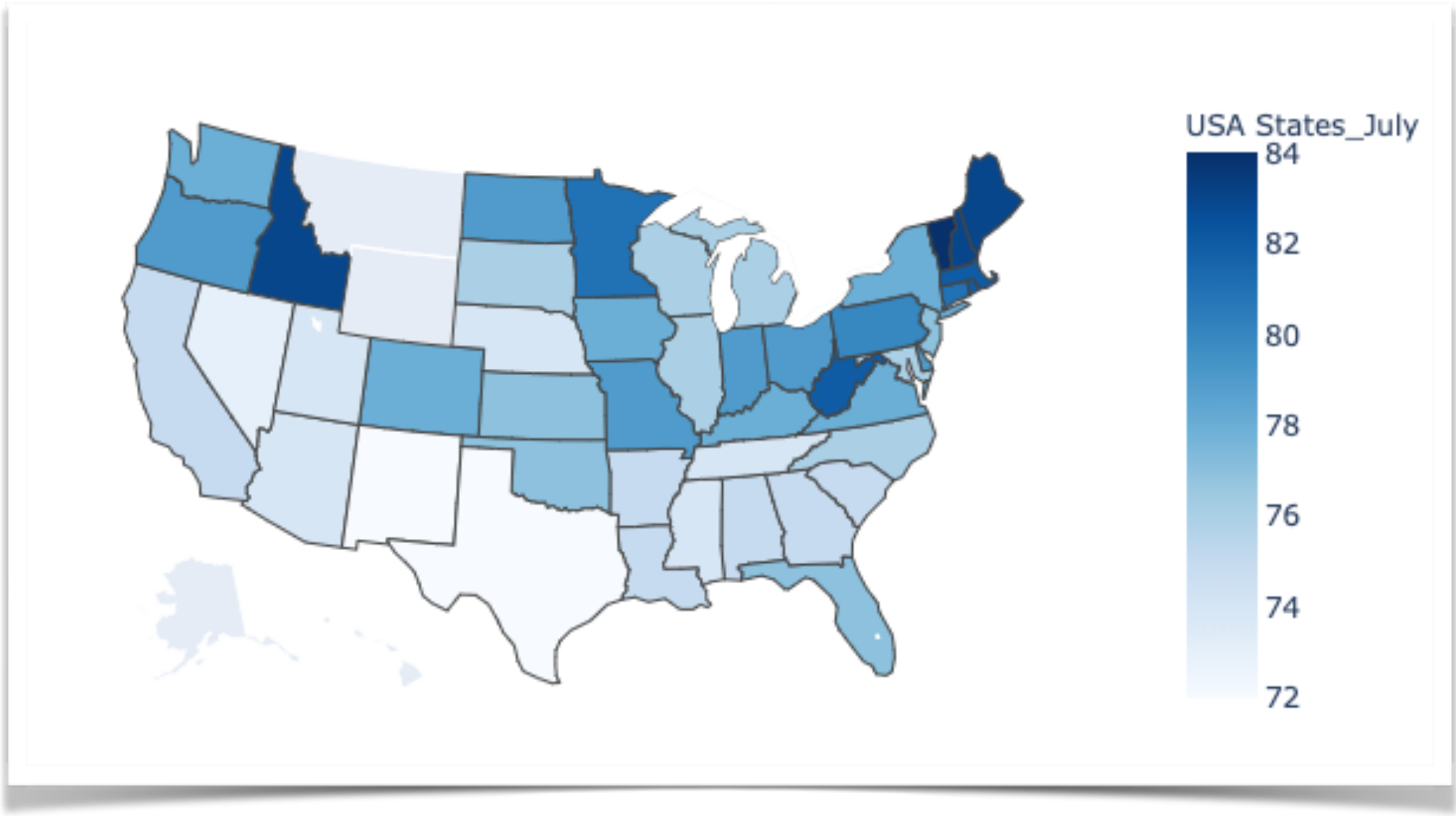

Figure 5: The distribution of average age group for each state w for the month of July

6. Assumptions and Limitations

The data assumes that all the deaths were reported on the cdc.gov website. There were few missing data, which was considered as no deaths, this can also be as “not reported”. The population of the states are different, this might effect the analysis but as the average deaths were considered, we assume that this has been taken into consideration. Also I consider that as the age group variation in the state is not taken into consideration, it might be that the state having higher average age deaths might be having more elderly population.

7. Conclusion

The analysis of data from the Centers for Disease Control and Prevention provides clear evidence of the toll that the novel coronavirus has taken on older Americans.

The analysis finds that states that have seen the largest share of COVID-19 deaths among people 65 and older include those that have had a disproportionate number of deaths in long-term care facilities. These states include Idaho (with 94% of deaths among those 65 and older), New Hampshire (92%), Massachusetts (90%), Rhode Island (90%), Minnesota (89%), Connecticut (89%), Pennsylvania (87%), Ohio (86%), Kentucky (84%), and Delaware (83%).

In this analysis for 11 weeks, with a 95% confidence interval we can say that the significance of a shift of the average age of 77 years for US over time, lies between 0.7 years to 10 years

8. References

1. CDC. Coronavirus disease 2019: cases in U.S. Atlanta, GA: US Department of Health and Human Services, CDC; 2020. https://www.cdc.gov/coronavirus/2019-ncov/cases-in-us.html

2. https://health.usnews.com/conditions/articles/why-young-people-should-care-about-covid-19