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

Neeharika Sinha, PhD

Aim : Analyze the COVID-19 mortality rate distributed across different age categories in the US population for the month of May, June and July

This might be a small step to answer the question of how to protect lives of our grandparents, adults and kids and stay safe at home

Data: 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, reported to NCHS by sex and age group for 52 States and Union Territories of the United States

Techniques: Statistical models and Data Visualization technique

Data Acquisition and Wrangling

- The weekly data for Provisional COVID-19 Death Counts by Sex, Age, and State was collected from the https://www.cdc.gov/website
- The date ranged from 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 file in xlsx format was downloaded and analyzed through IPython notebook
- The missing value (nan) was replaced by zero
- Figure 1. shows the information of the data frame
- The nan indicates either no value or the death was not reported. The nan values were replaced by '0' indicating no deaths

Figure 1 : Reading and data information

```
In [3]: df1=pd.read_excel("COVID-19_Death_Counts_5_6_2020.xlsx")
        df2=pd.read_excel("COVID-19_Death_Counts_5_13_2020.xlsx")
        df3=pd.read_excel("COVID-19_Death_Counts_5_28_2020.xlsx")
        df4=pd.read_excel("COVID-19_Death_Counts_6_3_2020.xlsx")
        df5=pd.read_excel("COVID-19_Death_Counts_6_10_2020.xlsx")
        df6=pd.read_excel("COVID-19_Death_Counts_6_17_2020.xlsx")
        df7=pd.read_excel("COVID-19_Death_Counts_6_24_2020.xlsx")
        df8=pd.read_excel("COVID-19_Death_Counts_7_1_2020.xlsx")
        df9=pd.read_excel("COVID-19_Death_Counts_7_8_2020.xlsx")
        df10=pd.read_excel("COVID-19_Death_Counts_7_22_2020.xlsx")
        df11=pd.read_excel("COVID-19_Death_Counts_7_29_2020.xlsx")

In [5]: 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
```

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

- 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 52 states and Union Territories including US overall rate for each week (total 11 week) through May to July was calculated

- Each of the range of age group was replaced by a mid value to get the average age as shown in Table 1.

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

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

- 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

- The red line shows the average age for US for each week

- Similar calculations was carried for all the 11 weeks starting from the month of May , June to July

Figure 1. Average death for age group ranging from 0 to >85 years for the week on May 6th 2020

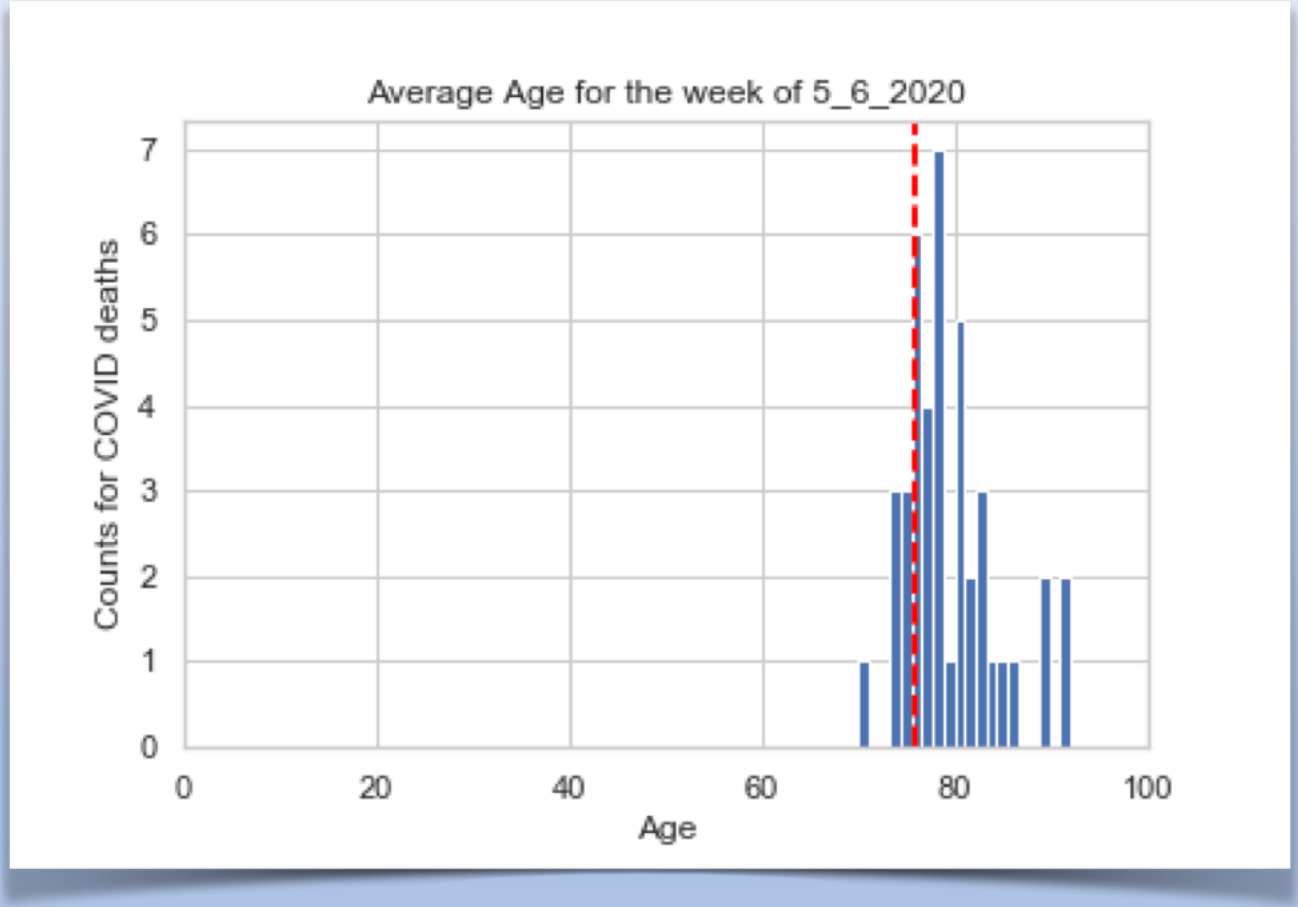
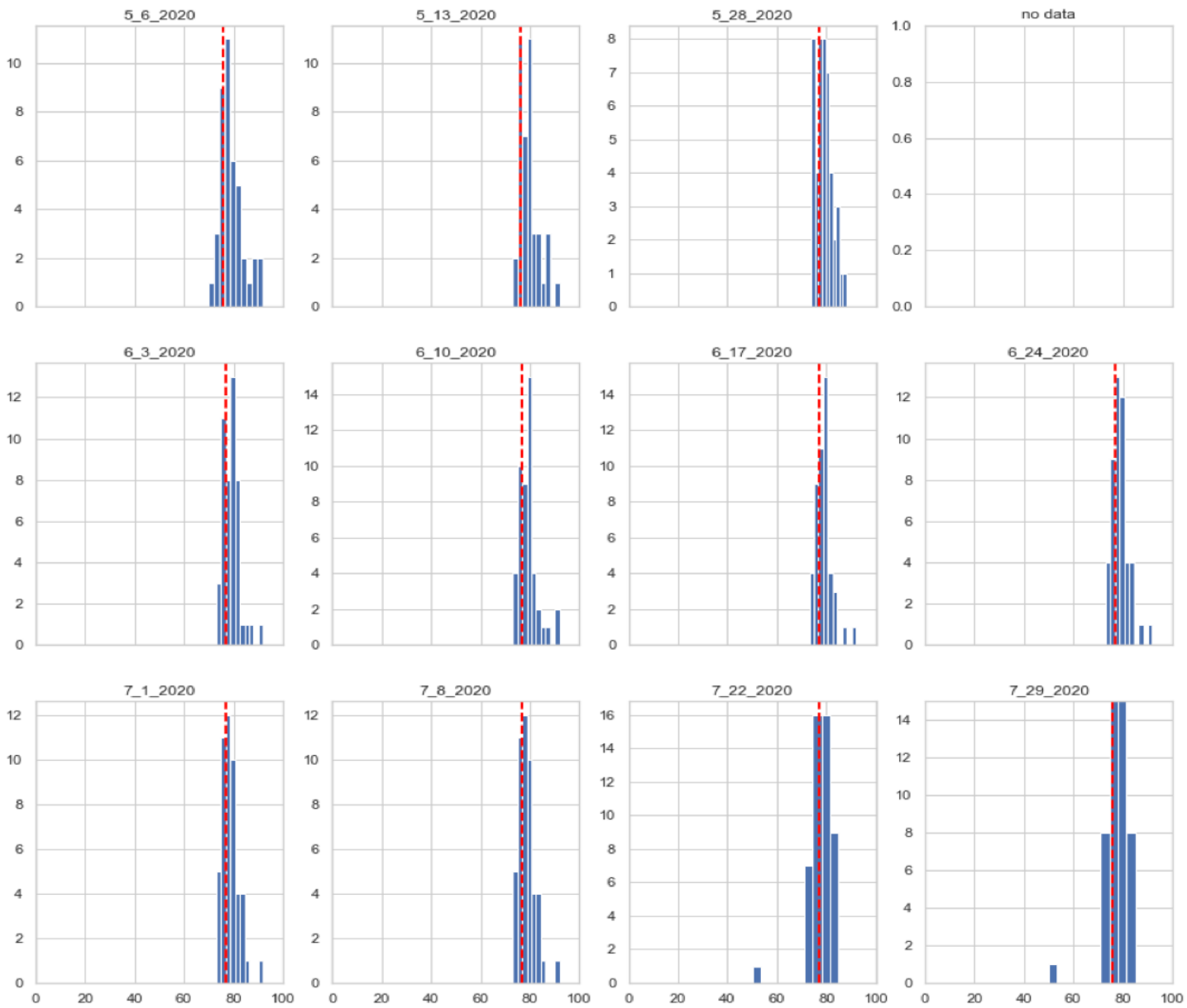


Figure 2 plots of average death for age group ranging from 0 to >85 years for 11 weeks starting from May 6th, 2020 to July 29th, 2020. The red line shows the average age for US for each week



- 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 by the virus

Statistical Model: T statistic

H_0 = 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)

Hα = the average death have changed through two weeks

Starting week	End week	p value
5/6/2020	5/13/2020	1(fails to give a conclusion)
	5/28/2020	0.66
	6/3/2020	0.57
	6/10/2020	0.71
	6/17/2020	0.43
	6/24/2020	0.38
	7/1/2020	0.25
	7/8/2020	0.25
	7/22/2020	0.06
	7/29/2020	0.05

Table2: The p value for hypothesis testing for the starting week to each of the next 10 weeks

- The t test failed for the immediate next week 5/13/2020, as the sample means in two groups are identical

- 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 , but the average age for the death remains the same

- 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

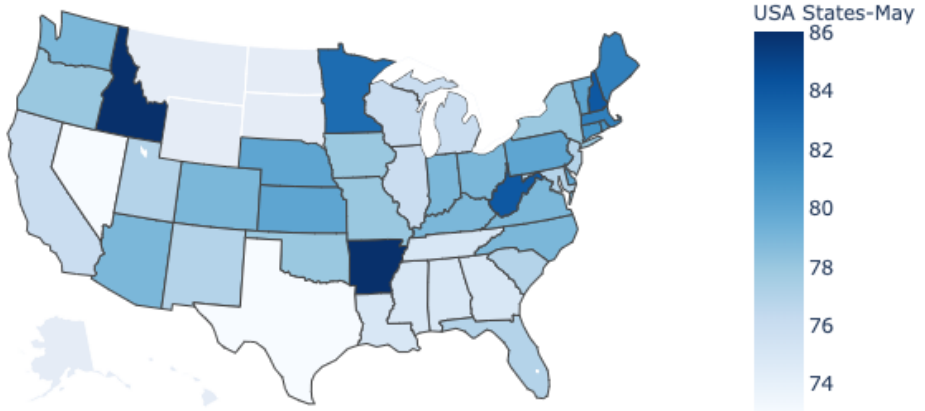
- Taking the difference of the average age group for all states and union territories, for consecutive weeks the variation in the average age through all states was analyzed

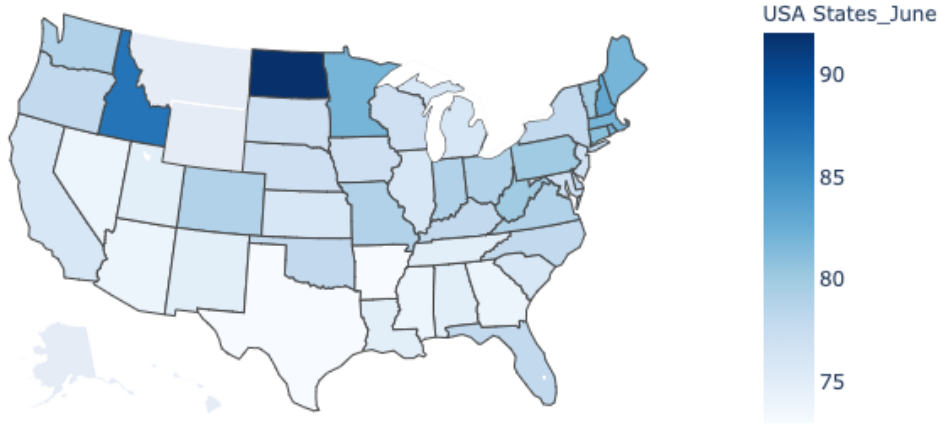
- With a 95% confidence interval we can say that the significance of a shift of the average age for US over time, in this 11 weeks lies between 0.7 years to 10 years

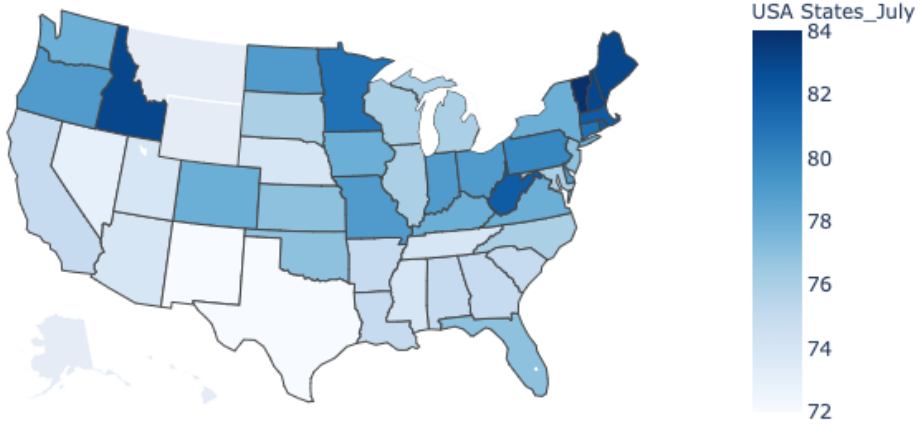
State Analysis

- The total deaths due COVID-19 for three separate months , May June and July was analyzed for 50 states of US
- 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 in Figure 3
- 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 in these dates, which needs to be confirmed
- 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

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







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.

Conclusions

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