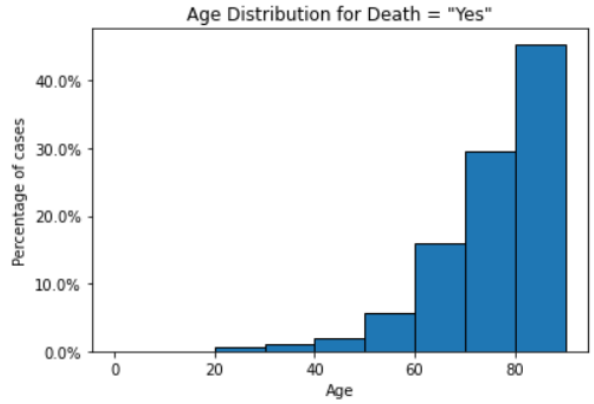
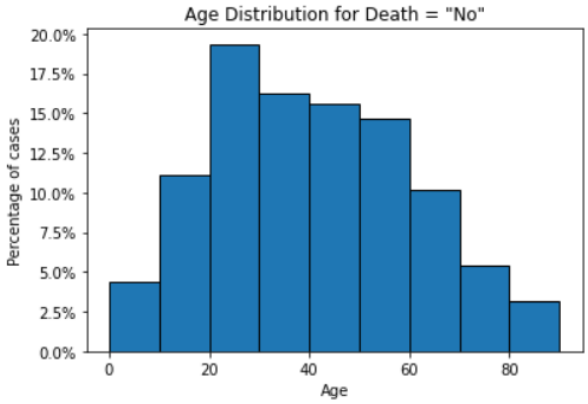
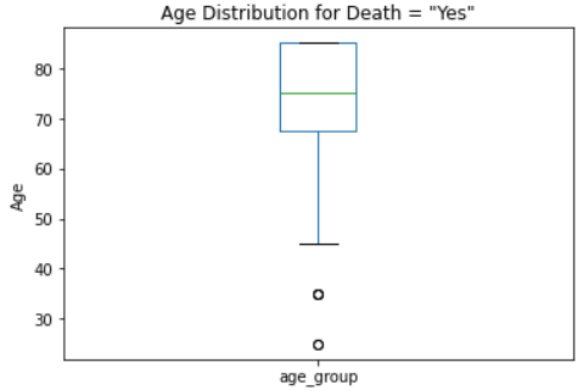
Part (3):

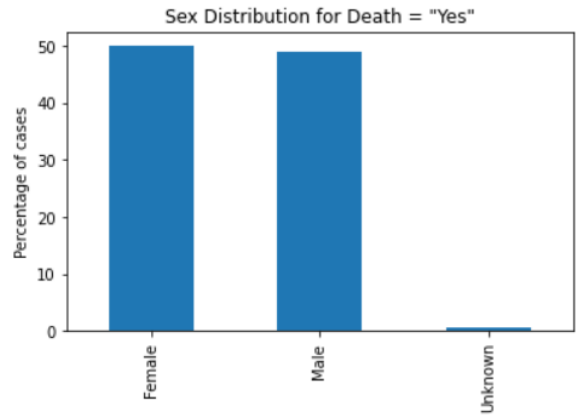
I plotted a histogram and boxplot of the age distributions for cases with death = “Yes” and death = ”No”. Clearly, the risk of death greatly increases with age. The age groups 70 – 79 and 80+ account for over 70% of deaths, but just over 10% of the sample. The histogram seems to show an exponential relationship between age and death.

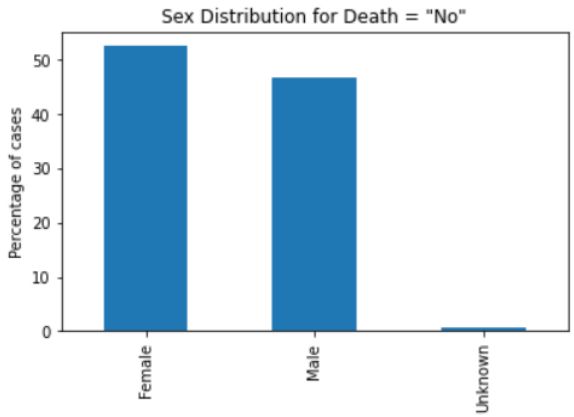


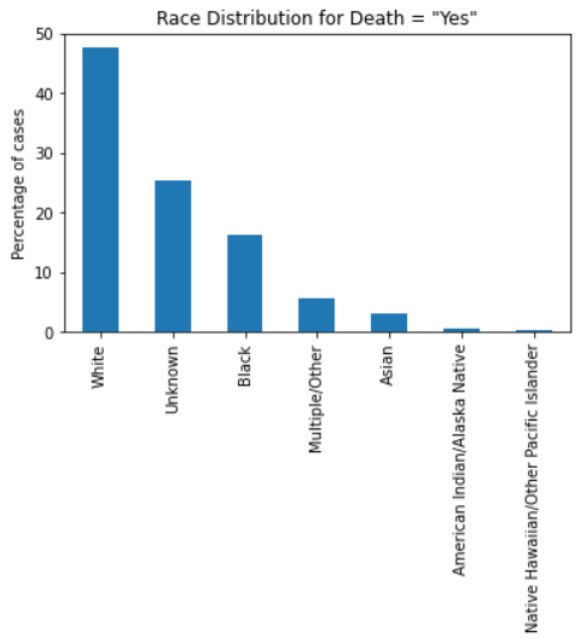


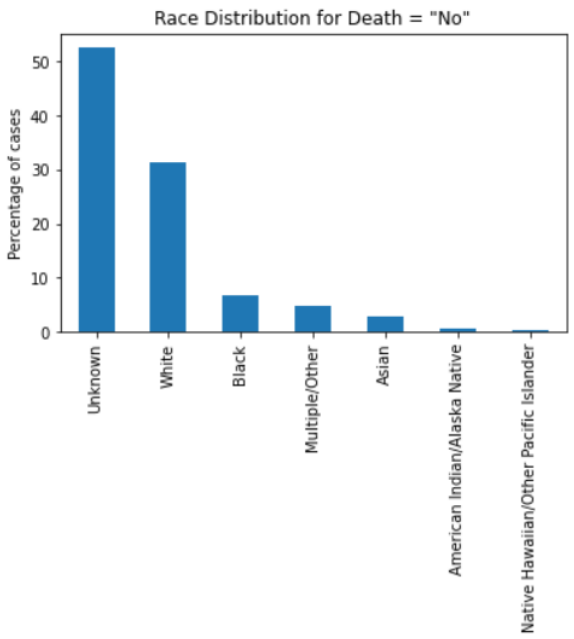


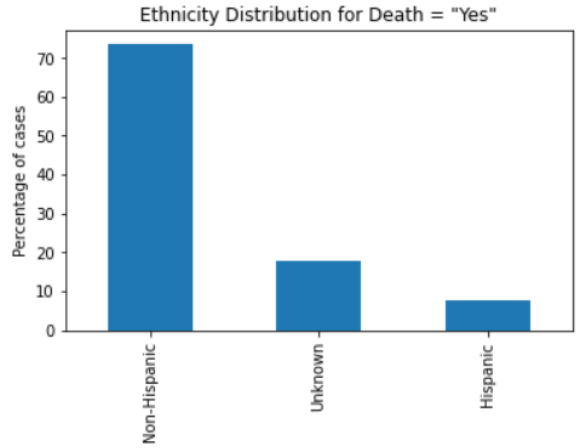
I also plotted a bar chart for the sex, race and ethnicity distributions with death = “Yes” and “No”.

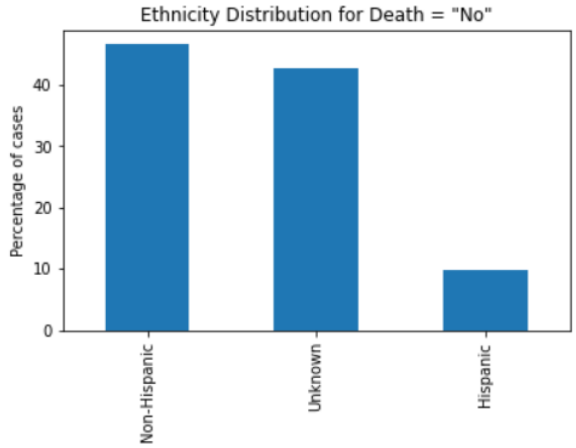












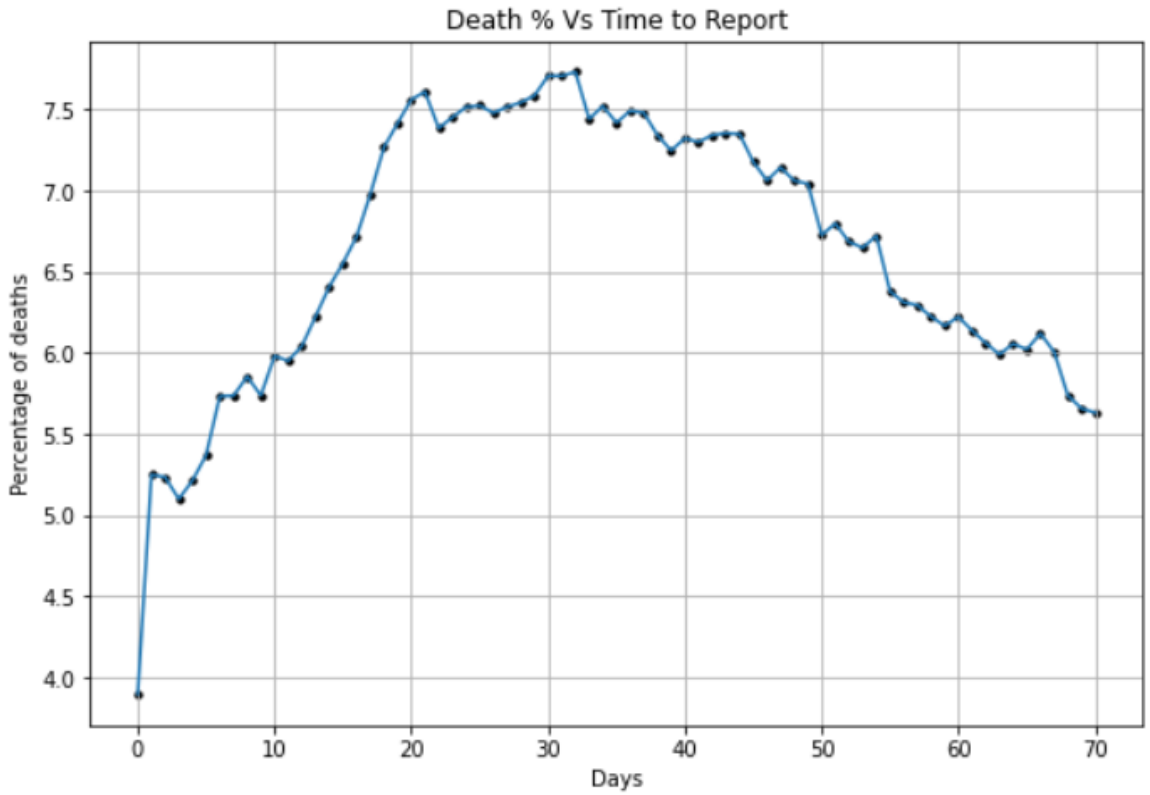
For sex, there were no significant differences between them.

For race, “White” increased from 31% of “No” cases, to 47% of “Yes” cases, and “Black” increases from 7% to 16%. However, “Unknown” drops from 52% to 25%. The changes in the other values are insignificant. It’s difficult to make use of this data with “Unknown” making up such a high %.

For ethnicity, “Non-Hispanic” increased from 47% of “No” cases to 74% of “Yes” cases, while “Unknown” drops from 43% to 18%. Again, this data isn’t really useful with such a high % of “Unknown” cases.

Part (4):

I added the feature “time to report”, by getting the difference between the earliest date and report date. I plotted the percentage of deaths Vs the time to report. Over 75% of cases were reported within 2 weeks. The percentage of deaths significantly increases for the first 3 to 4 weeks, and then steadily declines. Reporting as early as possible is very important to increase chances of survival.


Death % Vs Time to Report
Percentage of deaths
Days
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
0
10
20
30
40
50
60
70