**Steps/Strategies for cleaning data:**

1. Analyzed the research questions and identified the required columns for analysis of Excess Deaths data set
2. Reviewed the format of the column values in the data set to understand which command in R will be more effective to be used for this format
3. The summary of data was obtained using R to get an overall view of the data
4. Performed a check for missing data and found all missing data. Replaced all missing data with “NA”
5. Tried to figure out how can we deal missing value like deleting the observations, deleting the variable, prediction or imputation with mean/median/mode. By keen observation of the data, decision was taken to delete the observations as the incomplete data cannot be of much use for the analysis. There are enormous number of observations in the data set, where all the classifications are sufficiently represented in the data. For example, having the details of locality, age, cause of death without the metrics like Expected Deaths, Potentially Excess Deaths is of not much use for analysis
6. Checked the format of the data set to see which kind of command is needed to use in R and executed those commands in R Studio
7. We also found few duplicate values in the data set and deleted them using R command

## Issues with cleaning data:

## As this was the first time, all the team members are using R for data analysis, it took us time to get familiar with R. For example, creation of block for the code to run and closing the block before running the code

## Another issue was, we had the age range as 0-59, 0-64, 0-74 and so on but with the same locality, state and year. Initially, the team felt this might be redundant data. But later, when we looked at the total population values, the numbers kept increasing for each range. So, we finally reviewed the description of the data in the CDC website and understood that the age ranges were always kept at standard minimum value ‘0’. As there is no redundancy we didn’t delete any data.

## Deciding on which method to follow for the missing data was not an easy task. It consumed a lot of time too.