Data Wrangling

1. What kind of cleaning steps did you perform?

- First, I checked that the data was distributed on the basis of states inside the US. I want to see the figures of the US as a country, so I removed this column and grouped the data by date.
    - For mean, I aggregated the mean columns by grouping them by date and taking the mean of the mean columns.
    - For the upper uncertainty bound, I grouped by date and aggregated it as sum.
    - For the lower uncertainty bound, I grouped by date and aggregated the column as sum.

2. How did you deal with missing values and outliers, if any?

- Firstly, I checked the distribution of each of these columns and saw what shape each of them were taking.
- If the distribution was normal, I did missing value imputation by mean. That is, I replaced the missing values with the mean of the other values of the same column.
- If the distribution was skewed, I replaced them by median.
- Also, I used Imputation Using Multivariate Imputation by Chained Equation (MICE) and then saw which of the three methods were giving me better results and chose that method.

3. Were there outliers, and how did you handle them?

- I did the box-plot analysis of each of the columns and saw if the columns were having values which were lying outside of the whiskers.
- If yes, I took out the quantiles corresponding to each of these columns and replaced the outliers with the most optimum quantile values. Generally, it's 96 percentile value in case of outliers above upper whisker, and 4 percentile value in case of those below lower whisker.
- I also used z-scores to handle them. I replaced all values having z-scores above +3 or below -3 with +3 and -3 respectively.

4.What were the data wrangling techniques and tools you have used, including standard datasets and sources for the purpose.

I used the following:

- Fetching only columns of interest
- Datetime conversion
- Grouping column by date
- Setting index to date
- Isnull function for missing values
- Boxplot for outliers
- Outlier handling: Z-score based replacement