Covid-19 Project
Dhiman Sarkar
MS Business Analytics '21
Loyola Marymount University
Los Angeles, CA

Important notes with respect to the quality of data:

1. We will always notice a major drop in daily #'s in October since the dataset doesn't include the whole month, it only goes up to October 4th.
2. The dataset non-user-friendly representation because the positive tests and total test results have their data aggregated and it is added up. Therefore, a better representation is to add the date slider to display a more representative data for a specific day. --> Made new columns called daily death, daily negative, daily positive, and daily total results so that the pivot table does not return the sum of an aggregate.
3. Some States such as California do not provide some details such as recovered, Death Confirmed, Death Probable, etc. Hence, we cannot make up data for these columns during the data cleaning process.

I. Modified ASH (All States History) - New Sheet

1. Sort by State then Date.
2. Filter function to pick out state.
3. Moved the columns positive, total test results, and negative to the same area D-F column.
4. Made daily positive, total, and negative.
5. BOOLEAN: If state is not equal, then take current positive value, otherwise, subtract from previous -> Applies to daily positive daily negative daily total results.
6. Columns created: daily positive, daily total test, daily negative.
7. Turn off multi select when using slicer.
8. Sorted by month and sliced using states – Graph.
9. Looked at death column and noticed that it is aggregated as well so I will be creating a daily death column.

Observation: *While analyzing the data, it was noticed that there were minor problems such as wrong input. For example, state Arkansas (AK) on 2020-04-12 they reported 8038 total test results. However, the value decreased to 7830 on 2020-04-13.*

II. Modified ASH(Quality of Data)- New Sheet

1. Created a pivot table.
2. Included 'Quality of Data' as the rows along with the Months as Columns.
3. Daily total test results were used as the values so that the month returns an aggregate of the daily total test results.

4. A slicer was included to filter the data as per the states. User can pick single or multiple inputs.

III. Modified ASH (Death Ratio) – New Sheet

5. Created Columns Daily Deaths and Daily Positives.
2. Included Positive Ratio column by (Daily/Positive/Total Daily Results).
3. Included Death Ratio column by (Daily Deaths/Daily Positives).
4. Sort by descending order to find out the highest death rate month for each state.
5. The information can be filtered by states and displaying the relevant monthly data.
6. Did not use a graph to represent this because it's misrepresentative.
7. Tabular format would suffice the effective communication.

From this sheet, when you look at the data for NY specifically, you can see that after an upsurge in death rate in April and May, the death rate went down drastically. Furthermore, when you look at the sum of daily death for each month after May, we can see the numbers going down more.

Observation: NOTICED Death Increase is DAILY DEATH and Daily Death Can be removed.

IV. Hospitalize ratio (The sheet has been removed, but we had this our approach)

CLEANING UP:

Hospitalized and Hospitalized cumulative are the same (Remove Column)
Some States don't report data on hospitalized currently
Hospitalize increase is the daily increment in hospitalization
Hospitalized only counts covid 19 patients

Originally, we though that daily positive cases along with hospitalized increase to represent Hospitalized ratio is an okay step. However, we found that it is tricky because we can't use Hospitalized currently since we will aggregate data on existing patients
Hospitalized increase also takes death/patients who recovered into account
Furthermore, after observing the raw data, some patients recovered but weren't hospitalized! This indicates that not all patients who contracted covid-19 are hospitalized.

"Total number of people that are identified as recovered from COVID-19. States provide very disparate definitions on what constitutes a 'recovered'" - From Covid tracking

The way some of the hospitalization data is collected makes it very hard to create a hospitalize ratio.

V. Modified ASH (Positive Negative Ratio) – New Sheet

1. Columns included in this sheet were 'Month', 'Sum of Daily Positives', 'Sum of Daily Negatives', 'Sum of Daily Total Tests'
2. The data is sorted Monthly for the Selected State.
3. The output of the result is in the form of a pivot table.
4. Chose daily negative positive and total test results so that the pivot table returns a cumulative value for each month. If we did the original columns such as "Positives" the pivot table would return an aggregate of an aggregate which will mess up the data.
5. Along with the pivot table we also included a Multiple Bar Chart displaying Monthly Data of 'Sum of Daily Positive', 'Sum of Daily Negative', 'Sum of Daily Total Test'. The Covid numbers are reflected on the Y-axis whereas the Months are reflected on the X-axis.
6. Ratio is included outside the pivot table so that the pivot chart does not include the ratios (Includes columns 'Positive Ratio', 'Negative Ratio' and 'Total %').
7. The values in the table outside the pivot table is static. So the user has to copy the state-wise to get the updated 'ratios' and the corresponding '%'.

VI. Date Timeline (Sorting by States) – New Sheet

1. This one is good because it sums up the data up to a specific date instead of aggregating.
2. The timeline shows total # of cases from march up to a specific date.
3. It's showing the data at a specific point of time. This is why the aggregated data is not being summed up by the pivot table.