Milestone 1

Topic : Covid-19 related data preparation and visualization

Three Datasets :

1) **CSV File Data Source :**
   https://data.cdc.gov/NCHS/Conditions-contributing-to-deaths-involving-corona/hk9y-quqm

United_States_Conditi
ons_contributing_to_d

2) **Website data source (All Countries Covid-19 Testing and Confirmed Cases) Wikipedia:**
   https://en.wikipedia.org/wiki/COVID-19_testing#Testing_statistics_by_country

   Later on, the above source was changed to below source from Wikipedia (State by State COVID-19 statistics):
   https://en.wikipedia.org/wiki/Statistics_of_the_COVID-19_pandemic_in_the_United_States#State_by_state

3) **API :**

   a) COVID-19 Vaccine Distribution Allocations by Jurisdiction – Pfizer :
   https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg

   b) COVID-19 Vaccine Distribution Allocations by Jurisdiction – Moderna :
   https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws

Relationship existing between the Datasets or relationship to be created :

Dataset 1 : CSV file data source has Covid-19 related deaths numbers with respect to underlying conditions and causes in various states within US. So, The column to be used as common key with other two datasets would be the State column from this dataset. The data at country level within this column may need to be modified to reflect data as 'United States'. There are state level counts as well as aggregated counts entire US – national level.

Dataset 2 : Wikipedia website table data source has the details about the world wide country level counts of Covid-19 sample testing as well as confirmed cases along with data points like ***tests / millions*** and ***confirmed cases / millions***. I may have to modify the existing column name ***Location***, which consists of Country names to be able to use the data in conjunction with other dataset.

Dataset 3 : API data sources has information around the Vaccine distributions counts and plan among all the states in US at weekly level. These data sources have column name ***'Jurisdiction'*** which reflects the State name within US. I may need to modify the data column or create a new one to reflect the corresponding State Code like AZ for Arizona, NY for New York etc. Additionally, I may need to add a new column named Country code with data as 'US' or 'United States' in each row. This may be helpful to be able to join the data with other datasets.

Overall, looking to create Country code as a common relationship between all the datasets.

Interpretation of Data and Next steps for upcoming milestones :

Column Name	Description	Dataset Source
Data as of	Date of Data points collected for Analysis	**CSV File - Dataset 1**
Start week	First week-ending date of data period	
End Week	Last week-ending date of data period	
State	Jurisdiction of occurrence	
Condition	Condition contributing to deaths involving COVID-19	
ICD10_codes	ICD-10 code for condition	
Age Group	Age Group	
COVID-19 Deaths	COVID-19 Deaths	
Number of Mentions	Number of Mentions	
Flag	Counts less than 10 suppressed	

Location	Location as Country Name	**Website source - Dataset 2**
Date	Date of Data points collected for analysis	
Tested	Number Covid-19 tests performed	
Units	Units of actual tests performed (samples / cases)	
Confirmed (cases)	Number of Covid-19 positive confirmed cases	
%	Percentage of confirmed cases in relation to tests actual performed	
Tested / million people	Count of Tests performed per million people population in the country	
Confirmed / million people	Count of Covid-19 positive cases per million people population in the country	
Ref	Refrence links	

Jurisdiction	Jurisdiction	**API source - Dataset 3**
HHS Region	HHS Region	
Doses allocated week of 12/21	Doses allocated week of 12/21	
Second Dose Shipment (28 days later) week of 12/21	Second Dose Shipment (28 days later) week of 12/21	
Doses allocated week of 12/28	Doses allocated week of 12/28	

Second Dose Shipment (28 days later) week of 12/28	Second Dose Shipment (28 days later) week of 12/28	
Doses allocated for distribution week of 01/04	Doses allocated for distribution week of 01/04	
Second dose shipment for distribution (28 days later) week of 01/04	Second dose shipment for distribution (28 days later) week of 01/04	
Doses allocated for distribution week of 01/10	Doses allocated for distribution week of 01/10	
Second dose shipment for distribution (28 days later) week of 01/10	Second dose shipment for distribution (28 days later) week of 01/10	
Total Moderna Allocation "First Dose" Shipments	Total Allocation Moderna / Pfizer "First Dose" Shipments	
Total Allocation Moderna"Second Dose" Shipments	Total Allocation Moderna / Pfizer "Second Dose" Shipments	

The first dataset consists of the State wise data for Underlying conditions / causes for Covid-19 related deaths. I would like to visualize the statistics around these underlying causes of deaths and understand which ones are critical so that we can determine on deciding where more focused care / attention would be required for the Covid-19 patients.

The second dataset consists of country wise statistics of Covid-19 Tests performed and confirmed cases. It also provides information like Tests per million population and Confirmed cases per million population. These stats can help us compare the various countries in terms of Covid-19 cases. We can also try to visualize this data around total number of confirmed Covid-19 cases in comparison with total number of deaths counts from the First Dataset

The third API data source consists of US State wise Total number of vaccination doses for respective 1st doses and 2nd doses – one source for vaccines doses from Moderna and other one from Pfizer. This dataset also provides the total counts of the vaccines allocated for each of the state in US. I would like to use this data for comparison between the total counts of the combined vaccines allocated (from Moderna and Pfizer) in comparison with Covid-19 confirmed cases – this may throw some light on the allocation of vaccines doses is based on severely affected states in terms of Covid-19 deaths.

Overall, I would like to clean the data as needed by dropping off columns which may not be required in further data analysis / visualizations, adding new columns / updating the data elements in selected columns to maintain consistent relationships between various data sources using the country code and then create the planned visualizations.

Note : For API source, I am trying to gain access to API data available via CDC website. Exploring this option to see how to get the token and access to API. I may request help for gaining this CDC API access.