The entire project work – all of the Milestones have been a great opportunity to learn about the Data Preparation / Data Wrangling activities as a part of the Data Science project life cycle. We got to learn about the techniques like Discovering / Understanding the data, Cleaning and Enriching, Subsetting / Structuring / Formatting, Validating and Visualizing aspects.

To begin with, after selecting the topic of research, worked on identifying 3 different data sources with variety of data formats like csv / files, HTML web pages and API resource. More importantly we had to ensure that the 3 different data sources are related to each other via a common feature / column data, which would span across all of the three data sources. Or at least plan to select the data sources which, by adding additional feature(s), can be linked together. This common feature is greatly helpful for analyzing all the datasets together to form a meaningful and complete picture about the data and the story it may represent.

After some research, I selected below resources during Milestone 1.

1) csv data source from the CDC website – https://data.cdc.gov/NCHS/Conditions-contributing-to-deaths-involving-corona/hk9y-quqm

2) HTML source used for Milestone 5 – https://en.wikipedia.org/wiki/Statistics_of_the_COVID-19_pandemic_in_the_United_States#State_by_state

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

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

During the next three milestones – Milestone 2 through 4, we cleaned each of the three data sources using variety of techniques like Filling in / Replacing the missing values based on relevance, Filtering out or modifying the outliers data, Creating relevant Subsets from the datasets, renaming the features / columns with appropriate and short but meaningful names which are easy to understand and provide clear reflection of the information contained. Also, actions like duplicate removal and conversion / modifications of the relevant data records so as to be consistent across all the data sources and standards were completed. This helped greatly in preparing the data in clean readable format.

Once the data was properly formatted in each of the milestones 2 thru 4, we created some sample visualizations to understand the features and their relationships with each other, all within the individual milestone datasets. From these, I noticed and was able to confirm during milestone 2, that higher age groups were amongst the most affected in the population by the COVID-19 adversities and deaths. Also, Milestone 2 and 3 helped understand which US states were amongst the most affected with COVID-19 Cases and related deaths. During Milestone 4, I gathered the information around COVID-19 Vaccine Doses supplies allocated for each of the States and outlined which US States were getting higher number of Doses allocated.

For Milestone 5, I had to prepare the summarized version of the data from the milestone 2 through 4 to make it consistent across all data sources. We performed the database operations and loaded each datasets to individual SQL tables. We also learned about JOINing the three data sources to create Visualizations and highlight the relationships between various features from the three Datasets. It helped confirm that the Vaccine does allocations indeed were based on the highly affected states in terms of COVID-19 cases and related deaths. I have included the findings in the Jupyter notebook.