**Reflection**

In this assignment, I have analysis a real-world dataset named Covid 19 critical care online course which combined by 3 excel documents. Besides, each document has 8 worksheets. When I imported the data from excel documents to Python, I find that the Unnamed columns and rows with NaN values showed in python’s output data tables. The difference of input from excel and output from python is due to the different formats of the data. Therefore, I need to use loc function or drop the NaN values before I clean the dataset. According to the task assigned by the group, I will focus on analysis the Data run 1 for students. Then use another two datasets to comparative analysis in order to find the reason of changes and different relationships between different part of the dataset. Therefore, I also need to clean the rest of datasets.

In the dataset of run1, I find the data run1 has duplicate rows in the worksheet named visitors commenting on content. I check the course’s website that shows the course has 6 weeks but we get 11 weeks in the worksheet. Therefore, my idea has confirmed and I delete duplicate rows. About the missing data and messy code data in the rest of datasets, I choose drop the related rows because I cannot deduce or calculate these missing data or messy code data based on known data.

For the visualization part I tried scatter plot, pie plot, distribution plot, heatmap, linear regression, violin plot to analyze the data in order to discover potential relationships through the data. Also, I tried to draw a world map and use the map to show different nationalities of people‘s enrolments number of the course. When you move the mouse from one point to the other point, the map will highline different countries and its related course’s registered number.

I realized that real-world data is usually incomplete and needs to be analyzed according to different circumstances. Even if the data has been processed once, we also can get a lot of information from this processed dataset. When datasets involve multiple documents, we need to observe the data first. Then find the relevance before analyzing it. Finally, I learned how to use python to analyze excel files with multiple worksheets. After doing this assignment, I am able to analyze a wider range of data forms and will not be limited to analysis the CSV formats of dataset.