Q3

Since this is a very large file, I would consider imputing with mean, median, or the most frequent values because it is fast. It would be computationally expensive to use KNN, etc. However, it is mainly based on how many time that I was given to do the job. If I have enough time, I would love to fill it in with KNN, or MICE training, etc to gain more accurate result.

I do not like to put random numbers because I think it would hurt more on machine learning performance than leaving it at blank, as well as it spends time on filling in the data.

Q4

It actually depends on the context. I can write some code during the day that generally applies to the files I get every Thursday. So that I can simply click run to process the data with only minor changes, and let the computer do the rest while I am sleeping. Or if I am on the night shift, this is my job and I am fine with completing this task at 2am.

Q5

I would like to introduce a statistician who is also my professor, Professor/Doctor Mine Dogucu. I like her because she cares about things that people may not think of. After studying for only a couple of years of data visualization, I have formed a solid idea about data that data can only be visualized. However, Professor Dogucu uses one class to teach us about data verbalization, data sonification, as well as color usage in data visualization for color-blind people. This is important for giving everyone the same opportunity to access and use information. And I have a lot of admiration for her being able to think about it, act on it, and spread it.