For this project I am using a CDC dataset where they surveyed old adults (50 years old or older) about their health problems across all lands (states and territories) owned by the USA. I am trying to see whether there is any interesting inter-state relationships I can dig up from this dataset as I know nothing about this dataset but the information it contains should prove to be very interesting. One reason I picked this dataset is because it really is the only biology-related dataset that I found that I can use. Everything else were either strictly bioinformatics data, which proved hard to pick a question and solve for it (and many of them involves high level coding where an in-depth knowledge of bioinformatics and more importantly, that particular branch of biology [the -onmics field] is needed to solve the problem these dataset presents).

With the CDC dataset, it is broad enough to allow me to perform my typical data wrangling, cleaning and any python coding on it. It is also big enough that I can explore and see what interesting data stories I can come up with. This way I can showcase to a biotechnology company that I possess not only the python coding skills needed to work on a large dataset: I also possess the data analysis technical know-how needed to look at the data and generate a story from it.