JP Meyer | [kxn4mn@virginia.edu](mailto:kxn4mn@virginia.edu) DS 2002

This assignment has shown that data projects often require a significant amount of understanding of the project at hand before even beginning the coding or data section. This assignment was purposefully broad, and we learned that there are a lot of ways to reach a “correct” solution. Instead of immediately diving in and trying to code the solution, my partner and I spent a long time talking through our initial plan forward. We learned that the best way to do this was to speak through the assignment first, in words, which helped us grasp the concept of the assignment. It was particularly helpful to conceptualize what the final product would look like and work backwards to uncover the step-by-step process we would need to use to reach that point. Still, there were certainly some challenges that arose along the way. For one, I certainly understand the difficulty when multiple people are working on one data project at once. Originally, I somewhat misunderstood the significance of something like Github, where multiple people can make updates to a project without having to send updated copies back and forth constantly. While we did not use a Github repository for this assignment, I have a much greater understanding as to why it would be helpful. Instead, we used Google Colab, which was good for collaboration as it would save our work from time to time. It made collaboration easy enough and was incredibly helpful for understanding the errors in the code that arose as we were working. In terms of the actual coding, it was easier than I thought to get the base product to work. With some relatively simple if-else statements, we were able to correctly execute the assignment to convert data files across multiple data types. Our challenges arose with trying to add some ingenuity to our work – it was clear that there were tons of solutions for the assignment and some would be much more user-friendly while others required a bit more action from the user. We originally wanted to make the absolute best final product, but due to time and knowledge constraints, we ran into lots of issues. We tried our best to research and brainstorm quality ideas, but in the end, we had to take some shortcuts. It resulted in a lower quality product with perhaps less creativity, but still was effective in completing the assignment. This revealed a key finding – it is often required to make a tradeoff between quality and resources (time, money, knowledge, etc.). It is usually up to the data scientist or superiors to determine what level of quality is sufficient based on what is given, and the best decision-makers usually will find a successful level to balance this trade-off.

The requirements of the project also required that we uncover a very helpful skill in data science – being able to convert file formats for datasets and otherwise. Data comes in many shapes and sizes, often not cleaned and not in the format requested. It is highly important that a data scientist does not get caught up in these issues and throws out potentially helpful data due to formatting issues. Instead, being able to convert from and to any given format is very helpful. This can be applied across all levels of data science, but one example I can think of is for file intake forms on websites. Anywhere where you can upload a file somewhere and use the contents of the file in some way needs to be able to take in a file (often times, of any type), convert it to some format where the user can make edits, and then output the file into the desired type requested by the user. Depending on the complexity of the file and what edits need to be made, this is a vital but complicated process.