Data Wrangle Report

by Athar Ezzeldin

January 2021

I could say watching the lessons with the instructors explain everything to me ,made me feel it would be fine. However, in fact they paved the way to deal with data, and made it feel like an easy process, as I didn't have to think about the questions to answer or insights about data .With their guidance I know from where I should start. The quizzes had a big role in testing If I grasped the new ideas and I could continue or I need to spend more time. At a time I was excited to start my journey in wrangling. Other times I feel frustrated, especially because my educational background and experience is not related to that field in any way and this was my first exposure to learn this and apply it.

At the beginning of the project I faced challenges, especially the one that related to gathering data using the twitter Api, and I couldn't lie if they didn't provide the code and how to complete the process maybe it would took more time from me, but I will try to give it time to learn about the code. It took me time also to get the approval . I sent emails to twitter and answered their emails to get the authorization to try this on my own. After the gathering, came the assessing step which for me was okay and applied a lot of the assessments(programmatic and visual assessments) I have learnt during the lessons.

The cleaning process was quite challenging and I referred to many resources like Stack overflow, Pandas documentation, Geeksforgeeks and youtube videos to refresh my memory or help me with solutions that could apply to the cenario I am working on .After cleaning and merging the three datasets together to be just one DataFarme. I iterated on the assessing and cleaning again during each step .

I have to address the quality dimensions when cleaning for quality to make sure that the data is complete, valid, and accurate and for sure before merging I want to make sure there will consistency that's why I changed the Id to tweet_id to match the name of other columns in the other two datasets and make sure that the datatype is the same. These dimensions actually helped me to guide my thoughts when assessing and cleaning data.the image_prediction table was challenging somehow as it has tidiness issues and I was concerned about how to solve and maintain the same data. Mainly I worked a lot on the archive database and it tooks most of my time.

I thought for a time that analysis and visualization will be the easiest part. Here I felt confused, I need to find a pattern and look for insights which really aren't as easy as I thought. You have to address specific questions or get specific insight to use visualization and depending on what you want to address you could choose the best type that could serve your purpose.

I haven't solved all the problems and get the perfect analysis but I learnt a lot during my journey to finish that project and later and I will continue to work on it and learn more