Final Thoughts (Wrangle Report)

I have spent the most time on this project. I thought It was going to be easy and identical to the second project I took earlier on. I quickly realized that I needed to put much more work on it. Already in the gathering phase, specifically querying Twitter's API, it was clear that I needed to go over the material again, take notes, and practice in my own Jupyter Notebook. Querying of the Twitter API for each tweet in the Twitter archive and saving it JSON in a text was a cool experience for me because it was my first time parsing data from the internet. I found the structure (define, code, test) proposed for the cleaning process useful, but in my case it was clear that I needed flexibility and iteration. I had to go back to cleaning more (and making Google my best friend), even until the last chart was produced. I believe sticking to the process is a must, especially when dealing with multiple sources of messy and untidy data. Although it was recommended to start by solving tidiness issues, in this case removing first the unnecessary rows with retweets and replies, followed by the organisation in tables of observational units was important. When cleaning the datasets, I felt like I was being able to put into practice many of the skills I have been learning over the last months. Extracting HTML contents from a tag within the column of a pandas dataframe using BeautifulSoup was a nice accomplishment. What I wasn't sure about was whether my final dataset made sense because having different row counts makes the result different as what you expect. The key for me was to keep tweets and anything related (favorites, tweet_id etc) as the thresholds for all the data. When creating charts I used matplotlib and seaborn as I had been much accustomed to them in the past. The Highlight of my analysis and viz was adopting the time series analysis to compare retweet count and favorite counts. I read its documentation from Dataquest and plotted my visualizations on it and I felt really accomplished.