Wrangle Report

First of all, I import all the packages and modules that think i need to carry out the project. Some needed to be installed, like tweepy.

The datasets had different formats, so I opened them as a csv file to later work as a data-frame. It was somewhat complicated to use the twitter API but, in the end, I was able to use it to search through the tweet id of the first database for more information about each tweet, such as retweets and favorites.

After loading the 3 databases as data-frames I was able to internalize different aspects of them. On the one hand I realized that the WeRateDogs Twitter archive had many quality errors that had to be solved since it is our most important data base, through which we will join with the others.

There were many quality errors that were shown more frequently in columns that were not correctly formatted, such as 'Rating_numerator', 'Rating_denominator' and 'Tweet ID' in the WeRateDogs Twitter archive table. Another pattern that was repeated was the reply and retweet status columns that corresponded to duplicate tweets that we were not interested in, so i had to remove them.

Another quality error I found was that in the dog names column there were certain names that did not belong to dogs but were random parts of the tweets that were imported as names due to some error. All of these had a similar pattern, which was lowercase, so thanks to that it was somewhat easy to remove them.

Although there were more quality errors, the most difficult thing was finding tidiness errors. After a great research and joining the tables I was able to find two, on the one hand there were repeated columns in the 3 tables, such as 'Tweet ID' and in 2 'Timestamp' tables. On the other hand, the timestamp column was not in date format.

In conclusion, it was a very interesting project where I learned the modular spine of data analytics work, where wrangle data is such a complex process and that it requires a great dedication. The more we improve these steps, the deeper and more real the insights will be.