Wrangle report

In this project I put into practice the different phases of data wrangling: gathering, assessing and cleaning data. The first step was to collect the necessary data. For this I followed the project instructions to download the three necessary datasets twitter_archive_enhanced.csv, image_predictions.tsv and tweet_json.txt. For the json file I used directly the file shared in the classroom because for personal reasons I don't use twitter and I don't feel comfortable creating a user to access the data.

Then I started with the data assessment process. For this I focused on identifying data quality issues and tidiness issues. Quality issues refer to content issues such as duplicates, missing or incorrect data while tidiness issues refer to structural problems that can slow down the data cleaning, analysis, modeling and visualization process. In this phase I used both the visual approach and the programmatically approach. For the visual assessment I simply opened the datasets in pandas and reviewed them by scrolling. Here I was able to identify some initial issues such as missing data, some dog names that are not names and some strange rating values. In the programmatic assessment I used pandas functions to see more details of the datasets such as datatypes, maximum and minimum values that may represent data errors or outliers. With the help of these two types of data assessment I was able to detect the 8 quality issues and 2 tidiness issues that I documented at the end of the assessment phase.

Then I started the data cleaning phase which consisted of three steps: defining, coding and testing. Defining is the first step where the quality issue identified and documented in the assessment turns into cleaning tasks. Coding is the second step in which I execute the instruction in the defining step by using code. Testing is the last step in which we validate that the cleaning code worked. These same three steps were repeated for the 8 quality issues and the 2 tidiness issues. The result at the end of the data cleaning was the twitter_archive_master ready to start exploring ,finding trends and insights that allow us to understand the data and extract some knowledge from it.

With the master archive I was able to find the details of the tweet with the most retweets, which would not have been possible before as I had three different archives. We were also able to identify the top dog breeds and names as well as the distribution of dog ratings.