Data Wrangling Report

Intro

In this project I started from a single csv file containing a Twitter archive of WeRateDogs containing information regarding many of their tweets. The file was not well structured containing many mistakes from the text extraction process and was missing important informations like number of like, number of retweets etc. At this point I started the data wrangling process, dividing it in the following 3 main parts:

- Data Gathering
- Data Assessment
- Data Cleaning
- Combine all files

Data Gathering

Starting from data gathering, I first downloaded another file from Udacity containing URLs of the images of the dogs with their corresponding breed prediction performed by a neural network. As a second step I queried the Twitter API to gather all possible tweets present in the archive and saving them into a txt file as JSON objects. I then used this file to retrieve all relevant and missing information in the Twitter archive by paying attention to check if every single tweet was an original tweet or a retweet since I was interested only into original tweets.

Data Assessment

Going into the data assessment, I started with a programmatic assessment since the data was well structured and I noted down the following points:

- there are 163 retweets
- we can extract hashtags from the column "entities"
- sometimes the language is marked as not english but instead correspond to links and funny text, this basically means that all tweets are in english
- the column "id" should be named tweet_id like in the other files and should be of type string since we don't want to perform operations on it

Regarding the file containing pictures of dogs I found the following issues:

- "tweet_id" should be of type string since we don't want to performs operations on it
- we can compress the columns "img_num", "p1", "p1_conf", "p1_dog", "p2", "p2_conf", "p2_dog", "p3", "p3_conf", "p3_dog" into 2 useful columns: "breed" and "prediction_confidence" for tidiness

- I found 66 duplicated URLs of images
- there are 324 rows where none of the 3 prediction made from the neural network were predictions of dogs

I could have checked every single image url manually to double check the neural network prediction and/or filling up the missing values without any dog prediction, but I don't have enough expertise on this topic and it would also be out of scope for this project. I just want to mention that some prediction may be incorrect.

When I checked the Twitter archive file, I really saw there was a lot of work to do. It required me a couple of iterations of manual inspection and programmatic inspection and at the end here is what I found:

- also here "tweet_id" should be of type string
- there are some wrong values in the "rating_denominator" caused by bad text extraction
- ratings do not have all the time the same denominator
- there are many wrong values in the "rating_numerator" caused by bad text extraction, included a sing -5 marked instead as a 5
- "expanded_urls" is just a redundant information (https:// twitter .com/dog_rates/status/ + tweet_id)
- there are many wrong dog names called "a"
- all missing values in the columns "doggo", "floofer", "pupper", "puppo" and "name" have the string "None" instead of just being empty
- the columns "doggo", "floofer", "pupper", "puppo" should be combined into 1 called stage for tidiness

After all of these issues I should have also combined all these tables into a single tidy and clean file.

Data Cleaning

Data cleaning even if it was a lot of work, was done almost entirely programmatically. There was only 1 single case in which I had to clean the data manually, this was for the case of correcting a few denominator ratings values. This was caused by a non defined structure in the tweet, with some misleading information like i.e. this tweet:

"This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by the helicopter 10/10"

During the cleaning process I also noticed that 14 dogs had multiple stages like 'doggo' and 'pupper', this also had to be fixed by hand since some time the tweet referred to 2 dogs in the picture.

Combine all Files

After all this cleaning efford I combined all 3 files using the "tweet_id" column, discarded some empty columns, all rows non containing URLs of pictures and all retweets. This is the result of the final file:

Final File