Wrangling Report

Jaime Auger Esterio

Among the stages that the process had during its development are gathering the data, access the data and finally clean de data.

In the first stage, initially the necessary libraries were imported for the rest of the project (although throughout the project it was necessary to return to this point to include more libraries when necessary). The gathering data from a local csv file and the tweet image from a URL were easy and quick to code and execute. But the difficulty increased abruptly as soon as started working with the Tweeter API. It was a time-consuming process as it was necessary to wait for Tweeter's response to allow me to use their APIs. Next, the work with the Tweepy API was mainly guided by the tutorial they have in the API documentation. As product of this, 2331 JSON tweets were downloaded successfully and 25 had errors in their download.

In the second stage, the information was inspected using the info(), describe(), value_counts(), head(), isna(), duplicated () and Image () functions. In the table that I found the highest number of errors in the quality and order of the data was in the one of the file twitter_archive_enhanced.csv. The main issues consisted of wrong data types, wrong naming issues, retweets that were not required for, and the fact that the 'doggo', 'floofer', 'pupper' and 'puppo' columns should only be in one column. In the image_predictions table the main problem found was that there were 66 duplicate jpg_urls. Regarding the last table, this one did not present data quality problems.

Finally, each of the problems found in the previous stage was cleaned up. To begin, a copy of each of the databases was made and tidy issues were corrected first, since, according to what was read in the course, having tidy information facilitates the cleaning of quality issues. The first thing was to join the 3 tables into one using the function merge () on the tweet_id, to then immediately correct the problem of the 4 columns that should be only one. This was very helpful in understanding how the melt () function works. Among the other issues that were resolved, the one that required more work was separating the information contained in the source column into other new columns.

This is a list of all the issues cleaned during this process:

Quality

`twitter_archive_enchanced` table

- Incorrect data type format for 'timestamp'.

- All the not null entries of 'retweeted_status_id' are retweets and must be deleted.

- 'in_reply_to_user_id', 'in_reply_to_status_id', 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' columns are almost compose with missing values so they lack importance.

- Name 'a' is not a valid name. Could be better to change all the one letter names 'None'.

- The type of source and the source's url are in the same column.

- Correct the data type of rating_numerator and rating_denominator to admit decimals

- Create a new column that contains the obtained rating for each tweet.

`image_predictions` table

- 66 jpg_url are duplicated.

Tidiness

- The tweet_id column is repeated in all 3 tables, so it should be only one table

- Columns 'doggo', 'floofer', 'pupper' and 'puppo' from the twitter enhanced table should be only 1 variable.