Wrangle report

Wrangling started with **gathering** 3 data frames, by first, downloading the twitter archive which contains the basic tweet data like rating, dog name, and dog stage. Then I downloaded programmatically image_predictions.tsv file from Udacity's servers using requests library, last but not least I extracted the retweets count and favorites count for tweets in the twitter archive file using twitter API and using RegEx for extracting.

Then, I turned to **accessing** these 3 data frames, by first knowing the data types of each column is each file, getting the number of values of each column and checking for duplicates or any missing data in the 3 files, I did all this using pandas functions like info(), isnull(), duplicated(), sample() and value_counts(). Of course, I accessed the data visually too to discover the accuracy of the values and the tidiness too.

For tidiness, I found that -for example- counts should table and image_predictions table be part of the twitter_archive_enhanced table, (doggo, floofer, pupper, puppo) columns should be one column in twitter_archive_enhanced table and predictions should be in one column in image_predictions table.

And for quality there were non-original ratings (replies and retweets) and there were tweets for things other than dogs ...etc.

Finally, the cleaning was the most challenging part programmatically wise, first I made a copy of the 3 data frames to work with. Then I started to drop the unnecessary columns, and values (like the non-dog predictions) using methods like drop() and reset_index (). then I reextracted the ratings because it was inaccurate and removing the tweets which doesn’t include dogs, removing the duplicated URLs and extracting the null ones from the ID column.

Finally, I merged the 3 files together to make sure the data frame is tidy, but didn’t do that until I made the dog stage columns one clean column as well as the predictions columns.