WRANGLE REPORT

After making the three dataframes in the notebook I decided to look at them one at a time, in the following order: df1, df2 and at least the image_predictions.

The first thing I did was to use the info() function to see all the columns names, dtypes and how many non-null values it had. The first thing that catch my eye in the df1 was that the timestamp column wasn't a datetime type and the reply and retweet columns had a lot of null values, since most of the posts aren't of these two types. After I wanted to know which type of dog name is the most popular so after using the value_count on the name column the most popular wasn't a name but the letter "a".

That happened because the way the dogs have their name registered in the dataframe is that the text of the post is read, and after the word "is" the name of the dog is saved. But not all posts have the dogs named and only mention the breed so it becomes something like: "this dog is a husky". Those "a" names were changed to "None".

I wanted to look at the posts on twitter so when I tried to copy the link in the expanded_url table, I saw that most of the entries there had the link posted more than once separated by a comma. Some of the urls were also missing.

The doggo, floofer, pupper and puppo had most the values set to None and there was no real reason to have 4 columns to these values so I decided to put all in one column.

The projects rules states that posts that are retweets are not desired so those will be dropped.

In the df1 there are also posts that are labeled as reply so making a new column to label if that one row is a reply or a normal post can make it easy to visualize.

In the df2 as I was looking through the quoted_status entries I saw that one row had the quoted_status swaped in place of the retweeted_status.

This dataframe when looking at the null values, all the values in the geo, cordinates, place and contributes, are null making these columns worthless.

A lot of columns in the df2 have "str" in the name but are not of the object type.

In the possibly_sensitive and possibly_sensitive_appealable column all the values are the same but some of them are missing, so the missing values will be replaced by the same as the others.

The favorited column is a categorical one, where if the post is favorited the value becomes True, but some posts in the favorite_count column are above 0 but they are recorded as False in the favorited column.

As for the image_predictions dataframe, sometimes there is a dog in the image but the only way to confirm for how many entries this happens is to check one by one.