Wrangle Report

Introduction

A dataframe was created to merge :

1) Twitter_archive: dataset created by streaming the account WeRateDogs
2) Image prediction: variables dog type, confidence interval and true for three prediction attempts.
3) Favorite counts and retweet counts. : data obtained by the library Tweepy and a Twitter API

These three dataframes have been merged to get one full dataset. But as image predictions weren't available for every row, I've created a second dataset with Image Predictions.

A dataframe copy was created to preserve the original one.

One column doggo/floofer/pupper/puppo

I've created one column to say if there was any value in the "doggo","floofer","pupper" or "puppo" columns. If it wasn't the case, there was a "No_Nickname" value in this variable.

As I finished it, I've dropped the "doggo","floofer","pupper" and "puppo" columns from the database.

Dataframe cleaning : retweets and reply to

As not every row was a dog rate, I've cleaned rows with retweet status and reply to status from the dataset to get only the rating dogs' data.

Dataframe Cleaning : removing data without retweet and favorite counts

As the retweet and favorite counts are key insight, I've decided to remove all the data without these two variables.

Dataframe variable types: retweet counts, favorite counts, timestamp and img_num

Favorite counts, retweet counts, and img_num were in float instead of integral, so I've changed these types into the right ones.

As timestamp was in a string instead of a datetime type, I've changed the timestamp in a type more appropriate for dates.

Rating value

As the "rating_numerator" and the "rating_denominator" columns weren't usable, I've created a "rating_proportion" column, which divided the numerator by the denominator.

Source: URL address and anchoring text

As the source variable couldn't, I've decided to get two variables instead of this one:

1) URL address: for example http://twitter.com/download/iphone (so I've extracted this variable from the HTTP coding)
2) Anchoring text: for example: "Twitter for iPhone. This is the text which was contained the URL address".

Then I've dropped the "source" column.

Expanded URL:

To be able to use this variable, I've kept only one URL address for this variable. In most cases, there were duplicate URL addresses in the same column, so I didn't lose a lot of data.

Dogs Names

I've extracted names which began by "name is", "named", "meet" and "Meet". Doing this, I've reduced the "None" values to 676, and I've transformed these into Null values.

New image predictions database

As this was another type of data, I've created an image prediction database, and I've created a variable that considered only the first prediction from 3 possible ones if they contained a dog race. I've also transformed the image number into integral.

Then I've removed p1/p2/p3 variables from the main database and the number of images as well.