Wrangle Report

Student: Waldecyr Souza

Introduction

This document presents the efforts to gather, assess and clean the WeRateDogs twitter data.

Sources were obtained in several formats and the final objective was to prepare a single document containing all the useful information to answer the following questions:

What are the most frequent breeds?

What breeds receive more favorites?

What dogs stage receive more retweets?

For this, it was necessary to import several libraries, including Pandas, Numpy, Json, Matplotlib and Tweepy (this last one to access the data via Twitter API).

Development

1. df_breed

There were over 2k of tweets related images with the 3 best predictions regarding the breed of the dog. But the prediction of several rows was not related to dog breed. So we cleaned the data and in the end there were only one best prediction of the breed and the breed.

2. 2k_tweets

There were columns that did not not follow the pattern of the other WeRateDogs tweets; there were some denominator values different from 10 as well so it was necessary to look closely to it to fix then. To review data on quality and tidiness issues we put several columns in the pattern and dropped some inconsistent rows; we've turned some column labels into variables.
Besides that, there were two important column missing: retweet count and favorite count.

3. df_2k_twt_2_columns

In this project, we used the library Tweepy to query Twitter's API according to a previous list of tweet identification. After assessing it, it was only necessary to change the data type of the column 'tweet_id' to integer.

twitter_archive_master

After cleaning all the data above, we could merge all the files into one called twitter_archive_master using the key “twitter_id”.

Conclusion

The main data created was successfully used to answer the questions presented in the introduction of this document.