wrangle_report

September 18, 2022

1 WeRateDogs Twitter Archive - Wrangle Report

In this report I outline the wrangling efforts to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive.

1.1 Data Gathering

I gathered data from 3 sources, stored in separate files:

1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The favourite_count and retweet_count were extracted programmatically from this file.

I loaded the 3 raw data files into separate tables: archive, predictions and json_data.

1.2 Assessing & Cleaning Data

I started the assessment by viewing the information on the archive table first, I identified several quality and tidiness issues.

The columns that were not required for the project were dropped and this columns were 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' as they also had a lot of null rows.

The in_reply_to_status_id and in_reply_to_user_id column was converted to string data type as well as the timestamp column was converted to datetime data type.

The lowercase incorrect dog names have been dropped.

For this analysis retweets were not needed, all of the retweets were dropped.

The tweet id's that didn't have images were dropped

The gathered json dataset file was merged with twitter_Archive dataset into one table by doing this we dropped the tweets that didn't have matching tweet id's when gathering the json dataset.

The image_predictions tsv file was not clean since the was inconsistency in the dog breed names as some started with a uppercase letter and some lowercase. All dog breed names were changed to lowercase to have consistency.

Reduced four colums of dog types to one column in twitter_archive table so that (doggo, floofer, pupper and puppo) are not column headers and have one column as dog_class.

In [ ]: