WeRateDogs Twitter Archive Report

Introduction

The dataset that we will be wrangling, analyzing and visualizing is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage. The Twitter archive is great, but it only contains very basic tweet information so we are going to gather the data using variety of sources, assess the data and clean the data

Data Gathering

The data sources that I have utilized to gather data are below

Twitter-archive-enhanced.csv- This csv contains basic tweet data for all 5000+ of their tweets which only has ratings which brings down the dsize of this csv to 2356

Image predictions file-a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction

Additional Data via the Twitter API-twitter-archive-enhanced.csv does not have retweet count and favorite count which can be gathering by querying twitter API

Programmatic Assessment and Cleaning

While performing programmatic assessment I found several issues with Twitter archive data frame, for example, one of the common characteristics that can be observed is that the dog names start with capitalized letter, there are many

Dog names that does not start with a capitalized letter and most of them seemed to be invalid dog names. After the programmatic assessment, I have addressed all the issues that I found in assessment phase in the cleaning phase. Below are the

tidiness and quality issues that I have addressed

Tidiness Issues

1. There are 4 columns for 4 dog stages, I have melting everything into one column dog_type_value
2. Merge the df_tweet_data_extra into the df_twitterArchive_clean using inner join since everything is twitter daya.
3. Renaming columns corresponding to p1,p2 and p3 in the image predictions data frame

Quality Issues

1. Remove the columns which has more missing data and therefore not useful for analysis(in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp)
2. Remove the values that have numerator as 0(we have determined that there are 2 such values in our programmatic assessment)
3. Remove the values greater than 15 in the numerator(these are the values that are causing the mean to be more than what it should be)
4. Remove the values of the denominator which are not equal to 10
5. Change the data type of time stamp column to date time
6. Clean the source column and make it readable
7. Renaming incorrect dog names
8. Deleting rows which are not dogs from first_probability(p1)