DATA WRANGLING AND ANALYSIS

Introduction

The dataset that we will be wrangling (and 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.

Data Description

- **Enhanced Twitter Archive:** The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

- **Additional Data via the Twitter API:** Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? You're going to query Twitter's API to gather this valuable data.
- **Image Predictions File:** One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: 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 (numbered 1 to 4 since tweets can have up to four images).

Data Wrangling

- Data gathering,
- Assesing
- Cleaning
- Storage
- Visualizing and Analyzing Wranggled Data

DATA GATHERING

in this section, we will be gathering 3 different data from 3 sources;

- Reading in the provided twitter archive data
- Downloading the image predictions data from a server
- Querying the twitter API for addional twitter data

ASSESSING DATA

In this section, we assess the datasets we gathered to identify and highlight data quality and tidiness issues, this is also an excellent way to understand the data that we have acquired

- **Quality issues** which includes Completeness, Validity, Accuracy, Consistency : • unusual names for dogs like None,a,bo etc • numerator and denominator in ratings that are not according to rules • datatype for timestamp column in archive dataset • columns in archive like retweeted_status_id ,retweeted_status_user_id etc. • datatype for datetime column in tweet • user_favourites,user_followers are redundant columns in tweets dataset • missing values as number of rows not equal in all datasets
- **Tidiness issues** which includes structural issues : • stage variable in four columns: doggo, floofer, pupper, puppo • three different datasets for same data 'df_tweet' and 'df_image' and ‘df_archive'

DATA CLEANING

Cleaning for Messy Issues

in this stage, we will merge all three datasets into 1, and address all the tidiness issues we have identified in the assesing stage from assessing this data set, we are able to identify as couple of structuaral and data quality issues

- we can see we clearly have a lot of missing values,in columns; in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_user_id,retweeted_status_timestam (more than 80% of the data is missing) and the expanded_urls feature(having less than 5% missing in this case) -the name feature seems to be inconsistency in the names, that ought to be fixed
- we can also observe some messyness in the text columns, where we have links attached to the text, as well as the rating too. the column should ideally hold only text
- we can also observe that in the source feature, we have html tags enclosing the source. this ought not be the case

  Cleaning for Data Tidiness
- the last four columns indicate the dog type, all of that info can be represented in a single column, there also appears to be alot of None values, so roughly 60% of the values are None, they will be changed to Nan and later on dropped
- there appears to be redundant features in the dataset, for instance that are irrelevant for my analysis, these will be dropped
- the image number feature in the image predictions data set seems irrelevant to the analysis as such should be dropped

DATA STORING

storing our cleaned master dataset as a csv formate file for easy storage, other stoorage options include being saved in a database

Data Visualization and Analysis

Analyze and visualize your wrangled data in your wrangle data to derive insights and to communicate your