Wrangle Report

Introduction

This project aims to wrangle, analyze, visualize and gain insight into WeRateDogs tweet data. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog.

Dataset

This project uses 3 datasets:

1. Twitter archive enhanced: This is a CSV dataset that was provided by Udacity. It holds the WeRateDogs Twitter archive and provides most of the data needed for this project.
2. Image prediction dataset: This is a TSV dataset that results from a neural network classifying the images in Twitter archive data.
3. Supplementary Twitter API JSON dataset: This dataset is to be obtained from Twitter API using the Tweepy library. Due to difficulty in obtaining approval for Twitter API, the copy that was provided by Udacity was used for this study.

Data Assessment

The datasets were visually assessed by loading them into Microsoft excel and visual studio code as text files. The visual assessment was conducted to quickly develop a mental image of the structure of the dataset and identify quality and tidiness issues that can be uncovered by visual assessment. Next, the data frame was assessed programmatically by loading it into pandas data frame and executing functions to establish the shape of the dataset, column names, length, data types, row count and identify quality and tidiness issues. Some pandas functions used during assessment: `info`, `describe`, `isnull`, `sum`, `value_counts`, `head`, `sample`, `duplicated` e.t.c.

The following issues were identified:

Quality Issues

Tweet Archive Data

1. There are duplications from retweets or reply to an original dog rating tweet.
2. The project specifications specified that rating denominators are always 10, some rating_denominators values are not 10.
3. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp contains multiple NAN values.
4. The timestamp column is an object type, which is an incorrect type for the values that it stores.
5. The dog's name is not consistently capitalized. All the non capitalized dog names are either a single letter a, an, none or doesn't make sense for a dog's name.
6. Values for the source column contain redundant data not needed for this analysis.
7. Some observations are not original dog rating tweets, but retweets and replies.

Tweet API Data

1. There are 78 observations where the column in_reply_to_user_id values are set. These are all replies to tweets and not the original dog rating.
2. The columns in_reply_to_user_id, in_reply_to_status_id are all float64 types. They should be int64 types.

Image Prediction Data

1. The columns p1, p2, and p3 contain values that don't describe dog breeds. e.g web_site, laptop, notebook, wombat, street_sign e.t.c
2. Inconsistent capitalization of predicted dog breed.

Tidiness Issues

1. Merge the three (3) dataframes into a single dataframe since they all contain information about a single observation unit.
2. The columns doggo, floofer, pupper and puppo all relate to the same variable. A Tidy version of this dataset is one in which they are all in a single column.

Data cleaning

I had to make copies of all three data frames before cleaning the dataset. The data cleaning process followed three primary steps: Define, code and test. I had to follow logical order while cleaning the data because certain steps fixed other data issues; for instance, the image reply and retweet id columns on the enhanced data type had incorrect data types (float64 instead of int64). Dropping these columns took care of that too. Though a large part of the data cleaning process is programmatic, some processes required a visual examination of data. For instance, visually examining the tweet texts to determine the correct rating numerator. I am aware that I may not have cleaned every issue in the dataset, but I have cleaned enough to analyze the dataset.