WRANGLE REPORT

Udacity Data Analyst Nanodegree | Project #5 *(Analyze and Wrangle Data)*

INTRODUCTION

This project worked with data from the WeRateDogs Twitter archive, a Twitter account that features pictures of dogs alongside "ratings". Stages of data wrangling included gathering, assessing, and cleaning the data; this was then followed by data analysis and visualizations.

GATHERING

Data for this project came from three different sources.

1. **WeRateDogs Twitter archive**: all the tweets made by the WeRateDogs account up to August 2017. This was provided by Udacity and was downloadable as a .csv file.
2. **Tweet image predictions:** predictions of what dog breed (or other object) appeared in each image, according to a neural network. This was provided by Udacity in a .tsv file hosted on Udacity's servers and was downloaded programmatically.
3. **Retweet and favorite counts:** the number of retweet and favorite counts for each tweet in the Twitter archive. This was obtained from the Twitter API. Using Python's `Tweepy` library, I queried the Twitter API for each tweet's JSON data and stored it into a .txt file. The data was then read into a `pandas` DataFrame.

ASSESSMENT

After gathering each of the above pieces of data, the assessment stage required visual and programmatic inspection for quality and tidiness issues. I detected eight quality issues and two tidiness issues to be addressed:

Quality Issues

Twitter Archive Data

1. Dataset includes retweets and replies: values in `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, or `retweeted_status_user_id` indicate a retweet or reply (i.e., not the original rating). According to the Project Details, these will not be used in the analysis and therefore, these rows should be dropped.
2. Some denominators not equal to 10: while some of these are intentional, others are due to mistakenly taking the date (or another value) in the text as the rating. Since there are only a few cases, these can be remedied on an individual basis.
3. Incorrect `rating_numerator` values: some numerators were decimals in the original tweets and were read incorrectly.
4. Incorrect dog names: replace names in `name` that begin with a lowercase letter with null/NaN.
5. `source` column difficult to read: can remove HTML tag.
6. Incorrect datatypes: convert `timestamp` to datetime from object (string).

Image Predictions Data

7. Extraneous information in **image_predictions**: can create single `dog_breed` column.
8. Inconsistent capitalization/formatting of dog names in **image_predictions**
9. Unnecessary columns in **image_predictions**

Tidiness Issues

1. The columns `doggo`, `floofer`, `pupper`, and `puppo` in **twitter_archive** can be combined from 4 separate columns into 1 `dog_stage` column.

2. **image_predictions** and **tweets_df** can be joined with **twitter_archive** dataframe on the `tweet_id` column.

CLEANING

In this stage, I addressed the quality and tidiness issues identified in the previous stage.

Quality Issues

Twitter Archive Data

1. Eliminating retweets/replies was a simple matter of deleting all entries with values in the `in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,` or `retweeted_status_user_id` columns. These columns were then dropped.
2. There were, in total, 17 ratings with denominators not equal to 10 . Since this was a small number, I could examine each case individually. I found that the reason for these denominators was either due to 1.) taking a date or other numeric value as the denominator, 2.) aggregating the ratings of multiple dogs into one fraction, or 3.) a missing rating altogether. I dropped the entries with missing ratings, found the correct rating for the tweets that took the wrong number as the denominator, and scaled the fraction to a denominator of 10 for the ratings that had been aggregated.
3. All names beginning with a lowercase letter (indicating an incorrect name) were replaced with “None”.
4. The HTML tag was removed from the `source` column.
5. The `timestamp` column was converted to datetime.

Image Predictions Data

6. A single `dog_breed` column was created based on the prediction that was a confirmed dog breed with the highest confidence score.
7. Dog names were changed to have consistent formatting.
8. Unnecessary columns were then dropped.

Tidiness Issues

1. The columns `doggo`, `floofer`, `pupper`, and `puppo` in **twitter_archive** were combined from 4 separate columns into 1 `dog_stage` column.
2. All three DataFrames were joined together.

ANALYSIS

In the analysis and visualization section, I investigated 1.) average rating and 2.) popularity as measured by retweets and favorites for each dog breed (e.g., Golden retriever, Pomeranian) and each dog stage (e.g., puppo, doggo). Insights and conclusions from this analysis are discussed in `act_report.pdf.`