Wrangle Report for WeRateDogs Twitter Feed

By
Michelle Petersen

Overview

This report describes the process of Gathering, Assessing, Cleaning, Storing, Visualizing and Reporting on the WeRateDogs Twitter Account data. The title of the report begins with the word wrangling because data in the real-world comes from many places and is rarely clean.

Gather

Data for this project came from three sources.

1) The file “twitter-archive-enhanced.csv” that provided the basic tweet information. This file was downloaded manually.
2) The file “image-predictions.tsv”. It contains predictions from a neural network that can classify dog breeds based on an image. This file was downloaded programmatically from Udacity’s servers using the python requests library.
3) The third file was created from JSON data queried from the Twitter API using Twitter’s tweepy python library. The tweet ids in the “twitter-archive-enhanced.csv” file were used to query the API and the resulting JSON was saved to a row in the file named “tweet_json.txt”.

Each file was read into a separate data frame in a Jupyter Notebook to enable sssessment.

Assess

After gathering the various pieces of data I used the pandas and numby python libraries to assess the data for quality and tidiness issues.

Based on the project motivation the dataset and results should:

- Only contain original ratings (no retweets)
- Only contain tweets that have images
- Only contain tweets about dogs
- Focus on at least eight data quality issues and at least two tidiness issues

Data Quality Dimensions I looked for included Completeness, Validity, Accuracy, and Consistency. For each dataframe, I identified the changes that would need to be made based on both visual and programmatic assessment. The assessment and identification of issues were an iterative process. Many of the queries that I developed during the assessment were helpful during the clean phase.

In addition to the issues that needed to be fixed to meet the project motivation, I identified several data quality issues that must be fixed for my visualizations or analysis later on.

- The df_tweets table 'name' variable contains values that are not dog names such as ‘a' and 'an' 'unacceptable' and 'infuriating' that should be replaced with 'None'.
- Change rating_denominators to be 10.
- Change 'doggo', 'floofer', 'puppo', 'pupper' to be title case. Do not combine these columns since the for some rows there were entries in more than one column.
- Erroneous datatypes.
- Handle id and id_str correctly to comply with the Twitter documentation.
- Change column names to be consistent with Twitter API for consistency.
- Change dog breed names to capitalize each word and remove underscores and hyphens.

Tidiness Criteria:

- Each variable you measure should be in one column.
- Each different observation of that variable should be in a different row.
- There should be one table for each “kind” of a variable.
- Related tables should each include a column that allows them to be linked.

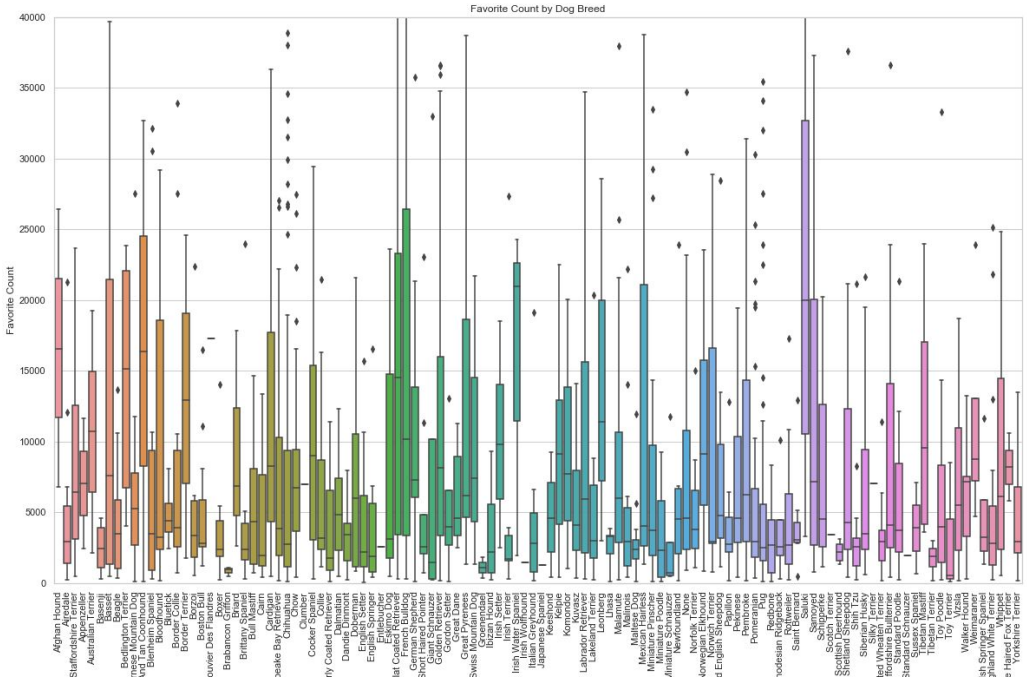
I combine df_tweets and df_tweet_additional_info and the image_predict tables so tweets and the data associated with a particular tweet id have one table. Three columns were duplicated across the df_tweets and df_tweet_additional_info tables in_reply_to_status_id, in_reply_to_user_id and source. I dropped the columns in df_tweets and keep the ones in the df_tweet_additional_info because these are from the extended tweet API.

Clean

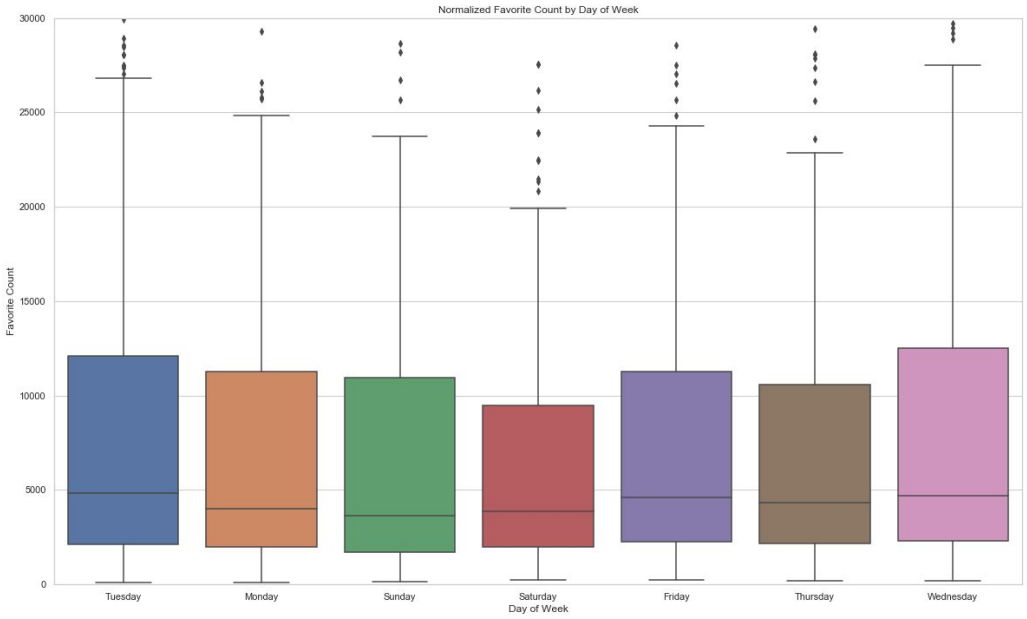
Cleaning the data was the next step and where I focused. Before starting the cleaning process, I made copies of each data frame so I could roll back to a last known good state. For each action, I followed the Define, Code, Test methodology. During the Code step, I took a measure of the data frame(s) to be modified before actions were taken. Some of the code wasn’t the most efficient but it was transparent and enable a test-driven process.

Visualize

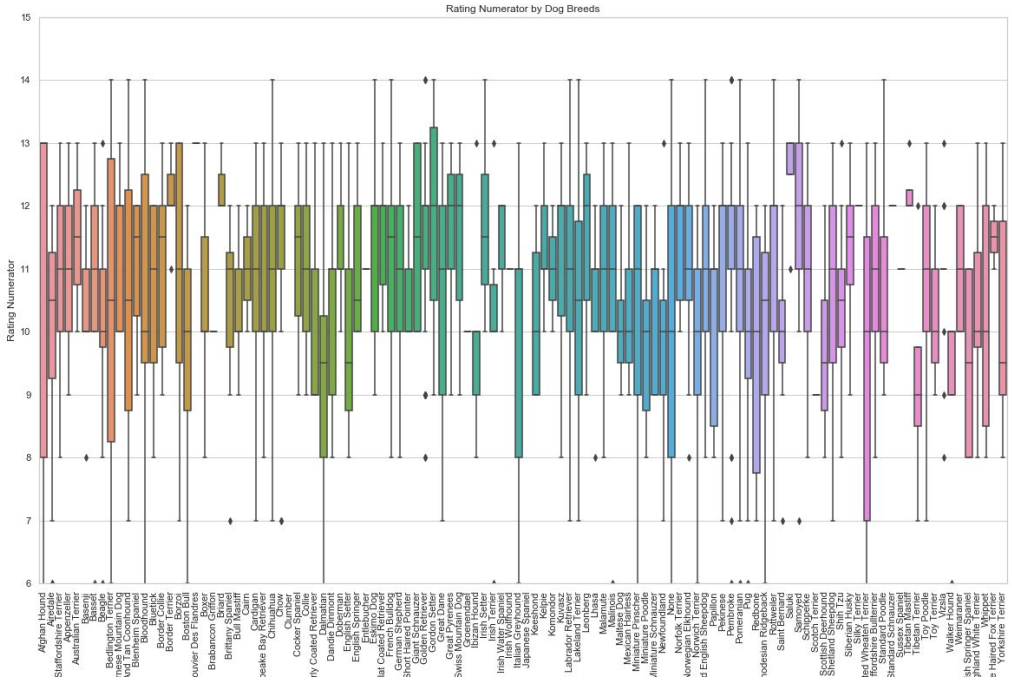
I created more visualizations than the project rubric required and enjoyed learning more about the seaborn python library.



The highest median favorite_count by Dog Breed is for the Saluki followed by the Black and Tan Coonhound and Afghan Hound.



The median favorite_count is higher on Tuesday, Wednesday, and Friday.



The rating_numerator for the Soft Coated Wheatlan Terrier had the lowest 25th percentile. The Redbone had the second lowest 25th percentile rating_numerator.

Summary

Wrangling is a necessary skill for any data analyst and I was thrilled to work on this project. I was surprised by some of the insights from visualizations - I expected to be looking at tweets primarily about golden retrievers. While working on this project, I received an opportunity to work on a small wrangling and visualization project for the California Master Gardeners Association at UC Davis. It is still a challenge for me to estimate how long it will take to gather, assess, clean, and visualize data because it is an iterative process and you don't know the questions you can ask or answer until you see the data.