Daniel Layfield

Data Wrangling Report

During this project, the requirements were to gather data from multiple sources and save them locally, assess the data, and fix tidiness and quality issues. Once this is completed, gather three insights, and create a visualization.

Gathering Data

Twitter Archive Enhanced – This data was provided as a .csv file and was read using the pandas pd.read_csv function.

Dog Breed – This data was available from a URL accessed with the requests library. The format was .tsv so using pandas pd.read_csv function sep='\t' was necessary to recognize the tab was what separated the data.

Twitter Data – This data was collected through Twitter (x)'s API. I had issues getting the data, so I used the instructor-provided data and the instructor-provided JSON data. Using this data, I gathered information such as Retweet Count, Favorite Count, and the full text of the tweet.

Assessing and Cleaning Data

Quality Issues

1. Irrelevant columns exist in both df_twitter and df_breed.
    a. In the Archive enhanced data and the Breed data, there were columns that I had no use for.

2. Confidence ratings below 51% are too low
    a. In the Breed data, if the confidence rating was too low for the first prediction, I dropped the record.

3. For incorrect data types, tweet_id needs to be an object, and the timestamp needs to be a DateTime object.

4. Ratings are too high or the denominator does not equal 10

    a. If the Numerator of the rating was more than 25 or if the Denominator of the rating was not 10 I dropped the record.

5. Retweets and @'s need to be removed so that only original posts are included.
    a. I only wanted to include original tweets and not retweets or replies.

6. Missing Images, any posts with no image need to be removed.
    a. If the data was missing an image then I dropped the record

7. Including only data before August 1, 2017.
    a. Data before August 1, 2017, was better not included per the instructions provided.

8. Posts that do not include the dog’s name should be removed
    a. I only wanted to include posts that contained the dog’s name.

Tidiness Issues

1. Multiple columns for dog stages
    a. These should be combined into a single column
2. Data is in many different data frames
    a. These should be combined into a single data frame for ease of access and centralization of the data.