WRANGLE REPORT

Gathering Data

The data were gathered from three sources. First, data were received as a comma separated file (CSV) from the lesson instructors. Then, using the request library, data on tweet image prediction were obtained from a uniform resource locator (URL). The dataset obtained was a tab-separated file which was saved to a local repository and read as a pandas data frame. In addition, a third dataset was queried from the Twitter application program interface (API) and stored as a flat file with data stored in the JavaScript Object Notation (JSON). Variables such as tweet_id, retweet_count, favorite_count, and retweeted were read into a dictionary and appended to an empty list. The appended list was converted into a pandas data frame.

Assessing Data

Visual and programmatic assessment methods were used to assess the following three datasets: twitter _archive_enhanced.csv, image_predictions.tsv, and tweet_json.txt. Nine quality issues and two tidiness issues were identified during the assessment stage. Quality issues included retweets and replies inside the dataset, missing data from in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp; Wrong datatypes with tweet_ids in all three gathered datasets, inconsistent sentence cases for dog names in image_predictions.tsv dataset, unlikely dog names in the twitter_archive_enhanced.csv dataset. Values with no entries in the dog stages were present in the image_predictions.tsv dataset. The source of the tweet column inside twitter_archive_enhanced.csv contains values embedded in html tags. Image predictions in image_predictions.tsv do not include all dogs or redundant columns. Tidiness issues included dog stages presented as multiple columns and tweet_id presented in multiple tables.

Cleaning

Identified issues in the assessment stage were defined, cleaned with codes, and tested to observe whether the overall cleaning goal was met. Prior to cleaning, copies of the original data frames were created. The data frame was merged using tweet _id as the key.

Storing Data

The combined data frames were saved as a CSV file.

Analyzing and Visualizing Data

The commonest source of tweets was from iPhones. Charlie and Lucy are the commonest dog names. The average rating of dog was 12.09. Less frequent dog names include Rose, Aubie, Kota, Leela, Glenn, Shelby, Sephie, Bonaparte, Wishes, Christopher. Two bar charts displaying the top ten names of dogs and the least ten names of dogs were generated.