Wrangle Report

WeRateDogs Data Analysis
Project: Wrangle and Analyze
Alyson Stiffel

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. For this project, I gathered data from WeRateDogs Twitter account and other sources in a variety of formats, assessed the quality and tidiness, then cleaned it.

Gathering Data

This project involved gathering data from three (3) different sources in three (3) different file formats and importing them into separate pandas DataFrames outlined below:

1. Read CSV using pandas and save as Dataframe `twitter_archive_enhanced`
2. Programmatically download TSV using Requests and save as Dataframe `image_predictions`
3. Query Twitter API for JSON data using Tweepy library and read into Dataframe `tweet-data`

First, the Twitter archive CSV file was read and saved into a DataFrame. Then, the image predictions TSV was downloaded programmatically from Udacity's servers using the Requests library. The image predictions detail what breed of dog is present in each tweet according to a neural network and is important for this analysis. Finally, using the Twitter API, I queried each tweets JSON data and stored the scraped tweets in a text file. This process is lengthy as the Twitter API response can be slow. The text file of JSON tweets is then imported into DataFrame.

Assessing Data

This next phase of this project is to assess the above DataFrames both visually and programmatically for quality and tidiness issues. The visual assessment involves scrolling through the data in Excel or using a function on `head` where we return all rows except the last *n* rows and/or print a random sample size. Whereas the programmatic assessment requires using code to view the data. For example, using pandas methods: `head`, `tail`, and `info`. To limit the scope of this assessment, at least two (2) tidiness issues and eight (8) quality issues are to be detected and documented.

Cleaning Data

Using pandas, I cleaned the tidiness and quality issues that were identified in the "Assessment" phase above.

Tidiness

I started the tidiness process by combining all three (3) DataFrames and saved them under a master CSV file. The second tidiness issue was combining the dog stage columns into a single column and renaming it for clarity. Then, the original dog stage columns were removed from the dataset as they are no longer useful.

Quality

The first point to address in the combined master DataFrame is duplicates. Any duplicated Tweet IDs as well as the retweets and reply columns need to be removed, in addition to any extraneous columns that

will not be needed for the analysis. I then continued by removing tweets with no image URLs and cleaned up the timestamp columns. This involved renaming and deleting a column as well as converting the remaining one to Datetime. Next, I addressed an issue with the `name` column where there were invalid or highly unlikely names. This was followed by cleaning the `source` column by removing the URL format and only saving the source text e.g. “Twitter for iPhone”.

I then converted all null values for the rating numerators and denominators. It was also important to pull and identify the incorrectly identified scores from the `text` column. This means the ratings were supplied in the text when they should have been included in their respective ratings columns. There were some additional outliers in the numerator values so I dropped any row with a numerator greater than fifteen (15) and ensured the datatype was converted to a float. In a similar vein, I decided to ensure all dominators were set to ten (10).

Next, I converted the remaining column datatypes. In the final steps of my cleaning, I made all image prediction strings lowercase and renamed the columns to be something more meaningful and easily understood e.g. “prediction_1” vs. “p1”.