Raymond Atherley

Wrangle Report

This report is an overview of the wrangling process for this project. This project required the gathering of twitter data for the WeRateDogs website from three different sources. The sources were from an on-hand CSV file, a TSV file programmatically downloaded from a website, and in json format through a twitter API.

The on-hand file was named "twitter_archive" and contained the majority of tweet and dog stage information. It contained the information found and extracted from a tweet and the tweet's metadata. This file as stated was directly provided to us and we simply downloaded it to our computer and uploaded it to the Workspace, as necessary.

The programmatically downloaded file was named "image_predictions". It contained three predictions of the breed of each dog using a neural network algorithm and based on the photo within the tweet. The range was from the best guess (p1) with a stated confidence value to worst guess (p3) and if the guess was indeed a dog (True) or not (False). In order to download this file, the requests package was imported. The specific URL where the data was located was provided. Code was written to request the data from the URL, to open a new file using a file handle and write the received data to the file in TSV format. This file was then uploaded and read into the Jupyter notebook.

The third dataset (JSON format) was named "twitter_API". It was based on the retweet count and favorite count for each tweet. The process to obtain this data was a lot more involved and required several steps as listed below.

1. I opened a new twitter account as I did not previously have a twitter account.
2. A developer account was requested and approved.
3. A twitter app was created, which provided the access credentials. These were a consumer key, access token and respective password/secret for both.
4. The tweepy package was imported.
5. Code was written to download the tweet data in JSON format using the access credentials and stored to a text file on my computer.
    a. the tweet ids from the twitter_archive file was used to request to status/information and using a for loop.
    b. an EXCEPT clause was used to record tweet ids not found in the API data
6. The text file was uploaded into the Workspace. Three specific fields (tweet id, retweet count, and favorite count) were read line by line into a user-created dictionary.
7. The dictionary was then used to create the data frame "twitter_API".

The three files were assessed for quality and tidiness issues. They were also cleaned as necessary. The issues ranged from incorrect ratings being extracted from the text column, and some rows having more than one dog stage. To correct this and condense the four dog stage columns to one, a FOR loop was used to check the value of each column and assign the value the single column. The number of records for each file also differed. The timestamp data type was corrected from string to datetime. The regular expression package (re) was used to extract the correct ratings and dog names from the text column. Inconsistent rows of ratings or describing more than one dog were removed as well as columns found not relevant. In addition, retweet rows were removed.