Report on Twitter Data Wrangle and Analysis – WeRateDogs

By Tania Couture

Introduction

WeRateDogs is a Twitter account that rates people's dogs and includes humorous comments and images. The goal of this project is to wrangle and clean WeRateDogs twitter data to create interesting and trustworthy analyses and visualizations.

Tools

Data wrangling, cleaning, and analysis was performed with Python utilizing the following libraries: requests, pandas, json, timeit, validators, re, numpy, matplotlib.pyplot, and tweepy.

Gathering the Data

- twitter_archive: The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. This .csv file was downloaded from Udacity via a link provided.
- image_predictions: A table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction generated by running the twitter archive through a neural network. This .tsv file is downloaded programatically using the requests library.
- tweet_json: Additional tweet data such as retweet count and favourite count not included in Twitter Archive. Lacking a twitter account and API this .txt file was downloaded from Udacity project though the code to query twitter is included.

Data Assessment and Cleaning

The 3 datasets were assessed visually and programmatically for quality and tidiness utilizing the define, code, and test method. The following issues found and cleaned:

Quality:

- tweets_archive:
    - not all tweets have url links to images - remove tweets without images
    - expanded urls have duplicated links separated by delimiter in same cell - remove duplicates
    - retweets created duplicate entries - remove any rows that are retweets
    - columns related to retweets are not needed - remove columns
    - many NaN values in table - the solutions above will fix this
    - text column contains unnecessary urls - remove urls
    - timestamp data type is Object - change to DateTime
    - not all dog names are names - change values that are not name to None

    - dog names have inconsistent capitalization - fixed when values that are not names are changed
    - rating denominators have large outliers - images with multiple dogs have large denominators - remove from dataset
    - some numerators have large outliers, ratings with decimals incorrect - remove novelty ratings and correct numerators that should include decimals
    -
- image_predictions:
    - inconsistent capitalization - replace underscores with spaces and change text to titlecase
    - duplicate image urls - delete duplicate images
    - columns p1, p2, and p3 contain items that are not dogs - remove any items that are not dogs then create one column for prediction and one for prediction confidence for the prediction with the greatest confidence rating

Tidiness:

- tweets_archive:
    - columns doggo, floofer, pupper, puppo can be combined into 1 column
- All 3 DataFrame should combined into 1

Analyses and Visualizations

Once the dataset was successfully wrangled, cleaned, and saved successful data analyses was able to be performed.

Conclusion

One challenge was familiarizing myself with the twitter and the wrangled data and format to discern what data was useful. The most difficult part of the cleaning process was finding more data issues as I was in the process of cleaning another issue.