Wrangle Report

Introduction Real-world data rarely comes clean and as a Data Analyst it's our job to gather, assess and clean the data to make it viable for analysis. Using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

Here, I am going to summarize my wrangling efforts done in the project notebook.

The dataset that I wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Data Gathering:

Data is gathered from three sources 'twitter_archive_enhanced.csv',' image_predictions.tsv' and 'tweet_json.txt.

- First data was collected from the "twitter-archive-enhanced.csv" file which was in the same directory in which project notebook was located. The csv file was imported into pandas dataframe. The dataframe was named "archive"
- Second data was extracted programmatically from a URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv.Python's request library was used to extract data from this URL. This file was imported as a dataframe in pandas using tab as the separator. The dataframe was named "image_pred".
- The third data was extracted from Twitter API using python's tweepy library. I needed to extract id, the favourites and retweet counts for each tweet. This data was then saved as a JSON file using UTF-8 encoding.

Data assessing:

Quality

- Faulty names
- Dataset contains retweets
- Unnecessary columns
- Tweets with no images
- multiple problems with dog ratings
- Difficult to read sources
- Incorrect datatypes (timestamp, source, dog stages, tweet_id, in_reply_to_status_id, in_reply_to_user_id)

Tidiness

- Dog stage variable in four columns: doggo, floofer, pupper, puppo
- Merge 'tweet_count' and 'image_pred' to 'archive'

Data Cleaning

- There are some faulty names in dataset
- Remove rows with 'retweeted_status_id' since we are interested in original tweets only.
- Remove unnecessary columns.
- Remove row where there are no images(expanded_urls)
- Combine dog stage columns (doggo, floofer, pupper, puppo) into one 'dog_stage' column
- Delete duplicated jpg_url in img_pred data frame
- Replace source links to string defining them.
- Change Dog Ratings
- Convert timestamp to datetime
- Condense Dog breed Column by choosing the one which is true at first as the first column has the highest percentage than the next two
- Change unusual or non-names to Unknown
- Join tweet_archive, image_prediction, tweet_json into one master dataset on tweetid.

Data Storage

- Saved the master dataframe to csv file “archive_master.csv”