Wrangle Report

1. Data Gathering Efforts

This project consists of 3 different datasets and the gathering efforts are given below:

- **Twitter Archive file**: twitter-archive-enhanced.csv was provided in the resource section and it was downloaded manually.
- **Image Prediction file**: he image_predictions.tsv was downloaded programmatically using requests library and URL information(https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv). This is scalable because it allows for setting dynamic filenames by extracting file names from the url and saving the content to specific file names. The mode='wb' allows for saving files individually but by setting mode='a', we can append content from numerous files into one file.
- **Twitter API and JSON file**: Downloading Twitter data via Twitter API and tweepy library was a lengthy and tiring process. At first the developer account for Twitter has been created and from there the consumer and access token/keys are found. In this project, a json parser is used by setting tweepy.API(parser=tweepy.parsers.JSONParser()) because twitter data is stored in json format and the json parser allows for converting to downloaded data in json format. Once the data is in json format, it is appended to tweet_json.txt file in the working directory. After the data is saved to the working directory, it is then read back into memory as a python list. Each element in the list is converted back into a json format and relevant data is extracted such as tweet id, favorites count and retweet count.

2. Data Accessing Issues

Quality Issues

completeness, validity, accuracy, consistency, variable(column), observation(row), unit(table)

- **twitter_archive dataframe**
  1. The expanded_url column, has some repeated multiple urls which is separated by a comma
  2. The name column has non name strings such as None, a, an
  3. Rating_denominator as high as 80
  4. The following variables should be integers instead of floats: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id
  5. Contains retweets
  6. The anchor text in source column is repeated numerous times

- **image_predictions dataframe**
  1. p1, p2,p3: upper and lower case mixed together
  2. p1, p2,p3: dash and underscore mixed in string eg. black-and-tan_coonhound
  3. Missing values when compared to twitter_archive dataframe

- **tweet_json dataframe**
  1. time_created should be a data/time objectlean
  2. dropping unnecessary columns

Tidiness Issues

- In twitter_archive dataframe: Variables called 'doggo', 'floofer', 'pupper', 'puppo' are different growth stages of a pet based on age, merge this in one column.
- Merge 3 datasets(Twitter_archive Data,Image_predictions Data, Tweet_json Data) into the final dataset.

3. Data Cleaning

The tiresome task is to clean the data. The format of timestamp in twitter archive data has been changed to datetime format for better refinement. The month, week and hour function in the datetime library can not be applied directly to a dataframe column and lambda function has been used for mapping. The following task has been performed in twitter_archive dataframe

1. format the timestamp to datetime format
2. add column for the month, weekday and hour the tweet was created
3. remove retweets
4. convert source to categorical value
5. remove duplicated multiple urls in expanded_urls variable
6. remove non name characters from name variable
7. dropping unnecessary columns
8. change the rating denominator to 10

The following task has been performed in image_predictions dataframe

1. dropping unnecessary columns i.e (‘img_num’)

The following task has been performed in json_tweets dataframe

1. dropping unnecessary columns i.e (‘full-text’, ‘time_created’)

Converting the ‘source’ column to categorical value really plays an important role just because we can foresee or tinker about the photo source of the dog’s picture. Unnecessary columns and duplicated values are also dropped.

4. Research Questions

The goal of working with these three data sets is to find some interesting aspects and reveal the frequent events. The following questions have been created for answering with the help of the final merged dataset.

1. Is the tweet that received the most favorites count also the tweet that was retweeted most?
2. What day of the week were most of the tweets created?
3. Which dog breed is most common ?
4. How do @WeRateDogs accounts write their posts? (DogCloud)