Data Wrangling Report

Gathering Data

Three different ways were used in Gathering the data :

1. Gathering the data from *"twitter_archive_enhanced.csv"* file, which was supplied with the project
2. Gathering the data from the url which was still supplied with the project as-well: *https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv*. This url contained 'image-predictions.tsv' which was then read and changed into a data-frame.
3. Lastly we used tweepy library to request the data needed from twitter for '*WeRateDogs*' . The file we captured was a json.txt file (tweet-json.txt) and then changed it to a data-frame containing those three columns ("twee_id", "retweet_counts" and "fav_counts")

Accessing Data

Visual Assessment Results

- **`twitter_arch` Visual Notes:**
  - `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id` and `retweeted_status_timestamp` columns most of them are NAN Values and they are not needed for our analysis
  - Some `rating_denominator` are more than 10 such as index (902)
  - The dog names are not all filled
  - Some `rating_numerator` are extreme

- **`twitter_img_prediction` Visual Notes:**
  - some prediction columns ain't needed we only need one for confidence Level and a second one for predictions

- **Can't see any visual problems in `twitter_api`**

Programmatic Assessment Notes

- **“twitter_arch” programmatic Notes:**
  - Data Type:
    - `tweet_id` column needs to be string instead of int64
    - `timestamp` column needs to be datetime instead of str but will be dropped
    - Inaccurate Data:
    - `expanded_urls` have missing data (2297 entery only)
    - Some `rating_denominator` are not out of 10
    - Some `rating_numerator` are extreme and not realistic
    - All Dog Classifications(`doggo`,`floofer`, `pupper`, `puppo`) null values are written 'None'
    - 181 retweeted tweets that needs to be erased from the data set before dropping the unwanted columns
  - Duplicated Data:
    - there are 137 duplicated `expanded_urls` which is an indication for the same tweet and therefore will be dropped
  - Only needed Columns in my analysis:
    - tweet_id
    - rating_numerator
    - rating_denominator
    - dog classification
    - expanded urls
- **“twitter_img_prediction” Programmatic Notes**
  - - Data Type:
    - tweet_id is int64 and should be changed to str
  - - Missing Data:
    - `tweet_id` got missing data (2075 entery) while expected to match the archieve (2356)
    - Tidiness of columns:
    - All dog ratings need to be only one column
- **`twitter_api`Assesment Notes:**
  - Data Types:
    - `tweet_id` is int64 needs to be str

Cleaning

twitter_arch Data Cleaning

1. Created a copy from twitter_arch to not make any changes in the original dataset.
2. Dropped the retweeted tweets.
3. Changed the data type of ‘tweet_id’ from int to str since no mathematical operation can be shall be done on IDs.
4. Replacing the ‘None’ values in all dog classification columns (doggo, floofer, etc…)
5. Created one more column called ‘dog_class’ and concatenated the three columns of the dog classifications together and renamed the ones whom have more than a classification. Finally dropped the three columns.
6. Removed the duplicated ‘expanded_urls’.
7. Removed the tweets that doesn’t had an empty ‘expanded_url’.
8. Removed the ‘rating_denominator’ that are not 10.
9. Fixing the rating_numerator by excluding the extreme entries: we will assume 20 is the max allowed range because such that it won't affect my analysis only 5 entries will be dropped.

twitter_img_prediction Data Cleaning

1. Created a copy from *twitter_img_prediction* to not make any changes in the original dataset.
2. Changed the data type of ‘tweet_id’ from int to str
3. Removed the duplicated jpg_urls
4. Reshaped the dataframe to have two extra columns one for dog type and the other was the confidence level.
5. Dropped the unwanted columns.

twitter_api Data Cleaning

1. Created a copy from *twitter_api* to not make any changes in the original dataset.
2. Changed the data type of ‘tweet_id’ from int to str

Merging the three Data Sets together using the tweet_id

1. Created df_combine1 that left merges (df_arch_clean) and (df_img_clean)
2. Dropped the null jpg_urls
3. Created df_main which merges (df_combine1) and (df_api_clean)
4. Stored the data frame to a csv as requested in a file called ‘twitter_archieve_master.csv’

Analyzing & Visualizations

1. Top and lowest rated breeds using mean value of rating_numerators divided by rating_denominator of each breed
   - Side Note: I thought mean value would be better than the sum of values to give a more reliable numerator rating.
2. Top and lowest retweeted Breeds by using the sum value of retweeted counts.
3. Top breeds that can be detected with high confidence using the img_prediction tool used.
4. Proportionality between the favorite counts and retweeted counts for the top 10 breeds
5. Checked if there is a relation between the rating and the retweet counts or the fav_counts but it appeared that there was no relation between rating and either of them.