WeRateDogs Udacity Wrangle & Analyze data

WeRateDogs is a twitter account where you can send a picture of your dog to and people can comment and give a rating to your dog.

Gathering data

First data was gathered to complete the project: 1. A 'twitter-archive-enhanced.csv' was loaded in. 2. The tweet image predictions were uploaded as well. 3. Tweet's retweet count and like count were uploaded using twitter API using the tweepy library.

Assessing data

Second the data was assessed by detecting 8 quality issues and 2 tidiness issues in the different data quality dimensions using the twitter archived data, images and twitter counts dataset:

8 Quality issues

Dimension 1 - Quality: Completeness

1. Tweet_id is an int (archive dataset)

2. Many missing data in the following columns: in_reply_to_user_id, in_reply_to_status_id, retweeted_status_user_id, retweeted_status_user_id, retweeted_status_timestamp (archive dataset)

Dimension 2 - Quality: Validity

3. Names of dog are sometimes just one letter or 'None' (archive dataset)

Dimension 3 - Quality: Accuracy

4. timestamp is an object (archive dataset)

Dimension 4 – Quality: Consistency

5. Difficult to interpret ratings (archive dataset)

6. Duplicated tweets

7. Various unnecessary image predition columns (image prediction dataset)

8. text column in archive contains both text and short link (archive dataset)

2 Tidiness issues

1. Each variable forms a column

2. Each observation forms a row

3. Each observation unit forms a table

1. The columns 'dogoo', 'floofer', 'pupper', 'puppo' all relate to the same variable (archive dataset)

2. The images and twitter_count dataset all relate to the same table (archive, images and twitter count dataset)

Clean the data

1. Define (see below)
2. Code (see wrangly_act.ipynb)
3. Test (see wrangly_act.ipynb)

Dimension 1 - Quality: Completeness

1. Tweet_id is an int (archive dataset)
   DEFINE: Change tweet_id from an integer to a string

2. Many missing data in the following columns: in_reply_to_user_id, in_reply_to_status_id, retweeted_status_user_id, retweeted_status_user_id, retweeted_status_timestamp (archive dataset)
   DEFINE: Remove unnecesarry columns in archive_clean table

Dimension 2 - Quality: Validity

3. Names of dog are sometimes just one letter or 'None' (archive dataset)
   DEFINE: Correcting dog naming issues

Dimension 3 - Quality: Accuracy

4. timestamp is an object (archive dataset)
   DEFINE: Timestamps to datetime format

Dimension 4 – Quality: Consistency

5. Difficult to interpret ratings (archive dataset)
   DEFINE: Standardize rating by calculating the value of the numerator divided by the denominator and save this in the column 'rating'

6. Duplicated tweets
   DEFINE: Keep rows with Nan retweeted, remove retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp

7. Text column in archive contains both text and short link (archive dataset)
   DEFINE: Create a function to remove links and apply it to achive_clean.text

8. Various unnecessary image predition columns (image prediction dataset)
   DEFINE: Replace unnecesarry image prediction columns by 1 column

Tidiness issues:

1. The columns 'dogoo', 'floofer', 'pupper', 'puppo' all relate to the same variable (archive dataset)
   DEFINE: Create one column for the various dog types - doggo, floofer, pupper, puppo

2. The images and twitter_count dataset all relate to the same table (archive, images and twitter count dataset)
   DEFINE: Merge clean version of archive, images and twitter count dataframes