WRANGLE REPORT

I had to 1. Gather data 2. Assess Data 3. Clean Data 4. Analyze data

DATA GATHERING

1. File name: twitter-archive-enhanced
   Format: csv
   Source: Udacity website.

2. Image Predictions File: Output from neural network
   File name: image_predictions
   Format: tsv
   Source:
   'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'

Variable	Meaning
tweet_id	the last part of the tweet URL after "status/"
p1	the algorithm's #1 prediction for the image in the tweet
p1_conf	how confident the algorithm is in its #1 prediction
p1_dog	whether or not the #1 prediction is a breed of dog
p2	the algorithm's #2 prediction for the image in the tweet
p2_conf	how confident the algorithm is in its #2 prediction
p2_dog	whether or not the #2 prediction is a breed of dog
p3	the algorithm's #3 prediction for the image in the tweet
p3_conf	how confident the algorithm is in its #3 prediction
P3_dog	whether or not the #3 prediction is a breed of dog

3. Data via the Twitter API
   File name: tweet_json
   Format: txt
   Source: Twitter API

DATA ASSESSING AND CLEANING

QUALITY

TWITTER ARCHIVE

1. Some tweet ids e.g '667509364010450944' has 'None', 'actually', 'a', 'an', 'getting, 'mad' etc for name.

I noticed the names that start with lowercase are the wrong names

2. Nulls represented as 'none' etc in doggo, floofer, pupper, puppo columns

3. Rating denominators are greater or less than '10'

4. id '835246439529840000' has Rating denominator of '0'

5. No dog stage for some dogs e.g '796116448414461957'

6. Most rows under 'in reply to status id' column are empty

7. Most rows under 'in reply to user id' column are empty

8. 'id', 'tweet_id' don't have same column names

9. 'retweeted_status_timestamp' and 'timestamp' have wrong datatypes

10. text column contains web links

IMAGE PREDICTIONS

1. Some image predictions have 3 'False'

TIDINESS

1. 'dog stage' variable is split into 4 columns - doggo, floofer, pupper, puppo

2. 'retweet count' and 'favorite count' columns are not in the twitter_archive

DATA ANALYSIS

After cleaning, data was ready was analysis and insights