Wrangle Report
WeRateDogs (Twitter Archive)

CONTENT

Gather Data

I gathered data from (3 sources), stored in separate files:

1. WeRateDogs (Twitter Enhanced archive), manually (downloaded from the Udacity servers).

2. The image predictions file, programmatically downloaded from the Udacity servers.

3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The favourite_count and retweet_count was extracted programmatically from this file. I loaded the 3 raw data files into separate tables: archive, predictions and json_data.

Assess Data & Clean Data

Quality 1. Twitter Archive: -
181 retweets
78 reply tweets
2297 tweets with expanded_urls
timestamp column is in string format
NOT a valid name 109 tweets
17 tweets with rating_denominator NOT equal to 10

Quality 2. image prediction: -
2075 image predictions, 281 less than

Quality 3. Json: -
TweetErrors

tidiness 1. archive: -
There are 4 columns for dog stages
Key Points indicates
rating_denominators

tidiness 2. image predictions: -
p1, p2 and p3 contain the same type of data

tidiness 3. json_data
combine json_data with archive table
