Wrangle report

Luyuan Zhang, August 2018

Data gathering:

Three separate information were gathered.
(1) archive data: downloaded from Udacity classroom.
(2) image_predictions data: gathered using requests library and the url for the dataset.
(3) updated_info data: more information of the tweets that are in archive. It is gathered through twitter API.

Data assessment:

After visually and programmatically inspect the 3 data set, I found following issues:

archive quality:

1. timestamps are strings, instead of datetime
2. replies and retweets
3. 13 tweets no longer exist
4. source values are difficult to read
5. multiple copies of url in same row
6. inaccurate rating_numerator and rating_denominator
7. multiple dog stages in single tweet
8. need one additional column "num_images"
9. inaccurate in dog names

archive tidiness:

1. external urls included in expanded_urls
2. dog stages spread into 4 columns

image_predictions quality:

1. duplicates of images and jpg_urls. This probably because some images are from reply or retweet

image_predictions tidiness:

1. only 1 predictions result is enough.

updated_info tidiness:

1. this dataframe belongs together with archive

Data cleaning:

Following steps were taken to clear the problems discovered during assessment:
1 merge archive and updated_info
2 convert timestamp to datetime
3 remove replies and retweets
4 remove non-existing tweets
5 extract clear source values
6 remove extra copies in expanded_urls in same row
7 reextract accurate ratings
8 melt 4 dog stages columns into 1
9 correct dog stages that have multiple values
10 correct dog names
11 remove duplicate images in image_predictions
12 drop p2 and p3 predictions in image_predictions
13 merge and save cleaned data

Data analysis were performed on the final clean data df.