Data Wrangling Report

Lirong Zhang

Data Gathering

First, three pieces of data were gathered: a twitter archive file in csv, a CNN image prediction file in tsv, and some extra twitter information for the twitter archives collected by using Twitter API. The twitter archive file twitter-archive-enhanced.csv is twitter archive from @dog_rates twitter user. The tweets are mainly about giving ratings for images of dogs of different breeds and stages sent by other twitter users. The tsv file image-predictions.tsv is a list of prediction of dog breeds based on the dog images by using a convoluted neural networks algorithm. The last file tweet-json.txt is a collection of favorite_count and retweet_count for (more or less) all the tweets in the first file.

Data Assessment

After data are gathered, they are being assessed. First all tables are loaded into pandas dataframes. Since the third table is an addition to the first table, so I merged the two together based on the tweet_id. Some of the main issues of the data are listed below.

QUALITY ISSUES

1. In merged dataframe, tweet_id and timestamp are of wrong data types.
2. In image_pred dataframe, tweet_id is of wrong data type.
3. The 'retweet_count', 'favorite_count' values are missing in two rows in twitter-archive-enhanced.
4. Some name entries are clearly incorrect, like 'a', 'the'.
5. 'None' is regarded as a valid name.
6. Extracted the wrong numbers as the rating.
7. Some tweets have no rating.
8. Some dogs appear to be in multiple stages.
9. Some images are not dogs.

TIDINESS ISSUES

1. Doggo, pupper, puppo, and floofer are all dog stages, should be in one column.
2. Some columns are irrelavant, so I removed the unwanted columns to not to be confusing.
3. All tables can be combined as one.

4. rating_numerator and rating_denominator should be combined into one single column for the rating.

Data Cleaning

This part turns out to be the most time-consuming section. I did not expect data cleaning to be not easy. I also learned that in real world problems, the data can be really messy and can take a huge amount of energy and time to clean them.