Data Wrangling Process

By Fangzhou Lin

1. Gathering

I gathered 3 pieces of data from different sources, in different formats and loaded them into 3 pandas dataframes:

- df_arc (enhanced WeRateDogs Twitter archive)
- df_pre (tweet image predictions)
- df_metrics (tweet ID, retweet count, favorite count for tweet_id in df_arc)

2. Assessing

2.1 Assess df_arc

I took a high level overview of the data. I observed extraneous columns and column values being headers. I also called info( ) to get a summary of all columns and found data type issues and null value issue.

Then I checked and found no duplicate in either record or tweet_id. But I found retweets data in df_arc. Missing value and incorrect input existed. I also found 23 records with denominator different from 10.

Last I checked the 4 columns of dog stages. I found multiple dog stages under the same tweet_id.

2.2 Assess df_pre

A data type issue was found. By comparing the number of unique tweet_id between df_pre and tweet_id in df_arc, I found out missing image data in some records of df_arc. Also, the data in df_pre should've been in the df_arc as 1 type of observation.

2.3 Assess df_metrics

A data type issue was found. Moreover, the data in df_metrics should've been in the df_arc as 1 type of observation.

3. Cleaning

3.1 Assess df_arc

I dropped extraneous columns and converted all the incorrect data types. Then I filtered rows with null values in retweet columns, in other words, getting all original posts without retweets. After that I dropped the three retweeted columns.

I decided to drop name column because pet names didn’t seem to offer much depth for my analysis. For multiple dog stages under same tweet_id I defined a function to unpivot dog stage columns, handle multiple dog stages and missing values.

I also dropped records with denominator different from 10.

3.2 Assess df_pre

I converted the incorrect data type.

3.3 Assess df_arc & df_pre

I filtered tweet_id in df_pre to only include those in df_pre. By doing this, I made sure that all tweets in df_arc were with image data.

I also joined the two dataframes and named it df_twitter.

3.4 Assess df_metrics

I converted the incorrect data type.

3.5 Assess ad_arc & df_metrics

I joined the two dataframes and named it df_twitter.

4. Datasets after Wrangling:

- df_twitter: tweet data including basic attributes and image predictions of breed

5. Storing

I stored df_twitter into a CSV file named twitter_archive_master.csv, and df_predic into a CSV file named breed_prediction.csv.