We Rate Dogs wrangling_report

Data wrangling have 3 steps to complete on datasets. They are

1. Data gathering
2. Data assessment
3. Data cleaning

Data has been gathered from different sources and loaded. Data assessment was conducted visually and programmatically against the quality and tidiness issues.

The following are issues noted:

Quality:-

twitter-archive-enhanced table:

1. No use of retweets columns.
2. Timestamp should be datetime instead of object
3. tweet_id should be in string instead of integer
4. in_reply_to_status_id, in_reply_to_user_id, expanded_urls have null values.
5. Name column have invalid names i.e 'a', 'an', 'the'
6. In rating_denominator column some values are more than 10.
7. we don’t need source column from tae_df table(most of the source information is repetitive).

image-predictions table:

8. p1, p2 & p3 have invalid dog names(other than dog breed names) like 'orange', 'bagel', 'toilet paper' and in p1_dog, p2_dog, p3_dog contains false , which means they are not useful.
9. We don’t need img_num column in ip_df table.

Tidiness:-

twitter-archive-enhanced table:

1. Dog stages should be in the same column.

Tweet-list table:

2. retweet_count and favorite_count columns should be in twitter-archive-enhanced table.

While assessing the data, I found that the dog name column has other than dog name. Which is important to clean. Some columns have wrong data type should be corrected because, it may cause difficulty while analyzing. They are null values for many variables. In addition , twitter-archive-enhanced table and tweets table are merge together and 4 columns of dog stages are merge together and formed single column named as stage to remove tidiness issues.

During data cleaning stage for each step cleaning procedure was documented in define.

Finally, the dataset was stored as .csv file.