WeRateDogs Twitter Archive – Wrangle Report

In this report, I document the wrangling process to gather and cleaning data which's used for analyze WeRateDogs Twitter Archive.

Data Gathering

The data used in this analysis is collected in three methods:

1. Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)
2. Use the Requests library to download the tweet image prediction (image_predictions.tsv)from Udacity. This data contains the prediction of dog breed by neutral network.
3. Use the Tweepy library to query additional data via the Twitter API (tweet_json.txt): Since the tweeter just update their developer account, my free account can't download the data. So I have to use the prepared data. This data contains the favorite count and retweet count.

Assess & cleaning

Though access these three data, I make clear of the consist of each table, each column presenting meaning and missing value, duplicated row in each table. Thus I summarize the quality and tidiness problems in the three table.

The *archive_df* , *tweets, predictions* tables have other 8 quality and 3 tidiness issues.

In accessing data, I found the retweet will produce duplicated text. So removing the duplicated row with retweet texts is needed. So I selected the rows whose the **retweeted_status_id, retweeted_user_id, retweeted_status_timestamp** columns are null. In this way, the duplicated text rows are deleted.

However, the in rely tweets just contain "@" in the text in order to reply or mention of friend. These rows provide meaningful context. Despite some missing values in the **in_reply_to_status_id** and **in_reply_to_user_id** columns, I deleted these two unnecessary columns in *archive_df* table.

In the dog "name" column in *archive_df* table, there are some invalid values like "a", "an", "the", "None", etc. I try to extract the dog from the tweeter text. By analysis the name appearance in text, The word "This is" and the uppercase in first letter in the extracted name become the indicators to extract the dog name.

At the end, dropping the row with Nan value in **name** column is also a necessary step.

After the investigate of rating score columns in in *archive_df* table, including denominator and numerator column, there exits such issue, such as the some values in denominator is not equal to 10. Since the text also contain the rating score, I extracted the numerator and denominator from the **text** column and change the string to float data type. Comparing values of the denominator and numerator column and the corresponding extracted denominator, I dropp the different rows. The last step is to drop the rows which the denominators are not equal to 10.

For the **timestamp** column in in *archive_df* table, I remove "+0000" and change it into datetime.

In the *archive_df* table, I drop the unnecessary **expaned_url** column.

To enhance the consistency and cleanliness of the **source** column in *archive_df* table, I focused on extracting the essential source information by using the string split function. The original source column contained both the URL and the actual source of the information, but for our analysis, we only needed the latter. By applying the string split function, I extracted the essential source part, which allowed us to have a more standardized and cleaner source column.

Regarding the quality issues in *prediction* table related to the predicted dog breed, I performed a comparison to identify the breed with the highest prediction confidence. The data from the image predictions contained multiple predictions for each tweet, along with corresponding confidence scores. To simplify the analysis and ensure accuracy, I created two new columns: one for the predicted dog breed with the highest confidence and another for its corresponding confidence score.

Considering the tidiness issue, I convert the four columns doggo, floofer, pupper, puppo, presenting the stage of the dog to one categorical column named dog_stage. The converted dog stage column posses some unknow dog stage and then replace them “None”. From the column value count, I find some row with multiple dog stages and try to extract the information from the text column. Due to no clue to extract the dog stage in text, I delete these rows with mixed dog stage.

Under the consideration of connivence of rating score, I convert the numerator and denominator into a factional form as a separate column named as rating score.

Since the tweets information spread across these three tables, I merged these three tables to combine an observational unit.