Data Wrangling

Prepared by: Muhammad Usman Siddiqui

Introduction:

This report discusses the data wrangling efforts on the data from the Twitter account WeRateDogs. The data is gathered individually from 3 different sources, then assessed, and then wrangled to be easier to interpret. Quality and tidiness issues of the data are identified and then addressed.

Identifying the Issues:

Firstly, the imported datasets are printed on the screen to visually identify any issue in the datasets. Then code is used to assess problems in the data.

The Issues and Their Solutions:

Firstly, copies of the datasets are made to make changes to them. The issues are numbered based on when they were discovered, but they are corrected in a different order.

Quality Issue 1: In the twitter_archive_enhanced.csv file, there are some incorrect names such as a, the, and an in the name column. These characters are identified and stored in a list, then replaced to null.

Quality Issue 3: The numerator values in the twitter_archive_enhanced.csv file contain some extreme values. In general, the numerator values should be a little above 10. To be on the safe side, values above 50 and below 7 are stored in a list and replaced by null. The authenticity of these extremes is questionable.

Quality Issue 4: The denominator values in the csv file should be 10. All values that are not 10 are identified in a list, and changed to null.

Quality Issue 2: In the doggo, floofer, pupper, and puppo columns, the null values are represented by ‘None.’ All None values are replaced by ‘’.

Tidiness issue 1: The doggo, floofer, pupper, and puppo are characteristic variables and are reduced to one column labelled dog_stage.

Quality Issue 11: In some rows, one dog is classified twice such as doggo and pupper for one dog. Not being able to check which stage is correct, these values are changed to, for example, doggo/pupper.

Quality Issue 5: There are non null values in the in_reply_to_status_id and the retweeted_status_id columns. These correspond to replies and retweets of the same dog, and not original tweets. Thus, rows with non null in these columns are removed.

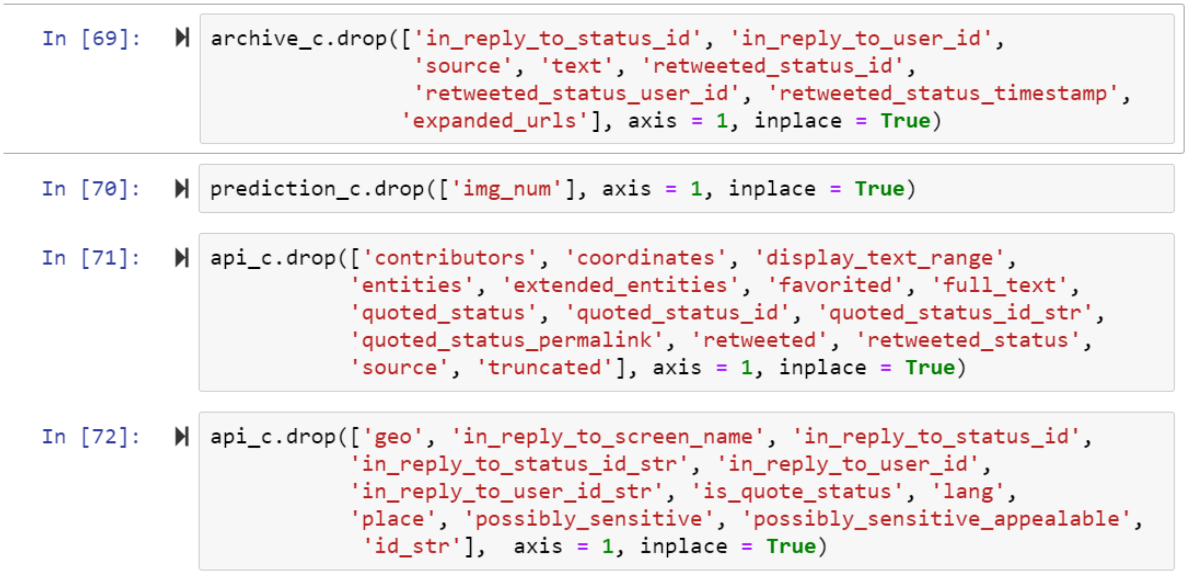
Quality Issue 9: The Timestamp column data type is changed from object to datetime.

Quality Issue 6: The values in the p1, p2, and p3 columns in the image_predictions.tsv file have non standardized capitalization and naming. All the letters are made uppercase, and all the '-' are replaced by '_' to make the naming standardized.

Quality Issue 8: The image jpg url is repeated for some dogs in the tsv file. Each dog has a unique url, and their repetition is a problem of duplicated rows. These duplicated rows are dropped.

Quality Issue 10: The non null values in the rating_numerator, rating_denominator, and expanded_urls columns do not match the rest of the data. This means some rows have incomplete data. Rows with null values in these columns are dropped.

Tidiness Issue 2: All the irrelevant columns are dropped as shown below.



```
In [69]:  archive_c.drop(['in_reply_to_status_id', 'in_reply_to_user_id',
                          'source', 'text', 'retweeted_status_id',
                          'retweeted_status_user_id', 'retweeted_status_timestamp',
                         'expanded_urls'], axis = 1, inplace = True)

In [70]:  prediction_c.drop(['img_num'], axis = 1, inplace = True)

In [71]:  api_c.drop(['contributors', 'coordinates', 'display_text_range',
                     'entities', 'extended_entities', 'favorited', 'full_text',
                     'quoted_status', 'quoted_status_id', 'quoted_status_id_str',
                     'quoted_status_permalink', 'retweeted', 'retweeted_status',
                     'source', 'truncated'], axis = 1, inplace = True)

In [72]:  api_c.drop(['geo', 'in_reply_to_screen_name', 'in_reply_to_status_id',
                     'in_reply_to_status_id_str', 'in_reply_to_user_id',
                     'in_reply_to_user_id_str', 'is_quote_status', 'lang',
                     'place', 'possibly_sensitive', 'possibly_sensitive_appealable',
                      'id_str'],  axis = 1, inplace = True)
```

Tidiness Issue 3: All three data sets elaborate on each other. Therefore, all three are combined into one data set based on the tweet_id.

Quality Issue 7: The number of rows of the csv file, the tsv file, and the API data are not the same. This means for some dogs, their image, their retweet count, and their favorite count is not present. However, this issue was resolved in the merging process.

Now, the data is ready for analysis.