Data Wrangling Report

Introduction

This project I will be focused on wrangling tweets archive and deliver interesting findings of Twitter user @dog_rate also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dog with humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage. This short report, I will be describing wrangling efforts which consists of gathering, assessing and cleaning data on WeRateDogs project.

Gathering Data

The data was gathered from three different sources:

1. twitter: twitter_archive_enhanced.csv is provided by Udacity and can be downloaded manually which contains 2356 row tweets archive
2. img_prediction: (image_prediction.tsv). The file contains prediction result of dog breeds. It was hosted on Udacity's server and was downloaded programmatically using the Request library with given URL
3. tweet_json.txt: retweet and favourite count were gathered and stored in txt file via Tweety library and Twitter's API

Assessing Data

The goal of assessing data is to improve quality and tidiness issue. Most of the data are messy, dirty and redundance and are not applicable for analysis.

Quality Issues

1. HTML entities found in 'source', can be fixed by extract important text to enhance readability.
2. Incorrect 'source' datatype. Convert 'source' datatype to categorical.
3. 'timestamp' can be converted to datetime format
4. Incorrect rating denominators, it should be 10
5. Some rating numerators are too large
6. We are going to assess and analyse "original tweets", no "retweets".
7. Null represented as 'None' in columns 'name'
8. Uncapitalized texts are incorrect in 'name' column. Example "a", "an", "the" etc
9. Image prediction contains data redundancy

Tidiness Issues

1. twitter's doggo, floofer, pupper, and puppo columns should be merged into one column
2. All three files have tweet_id column, which can be merge into one dataframe.
3. Dropping unnecessary column that are not useful for analysis.

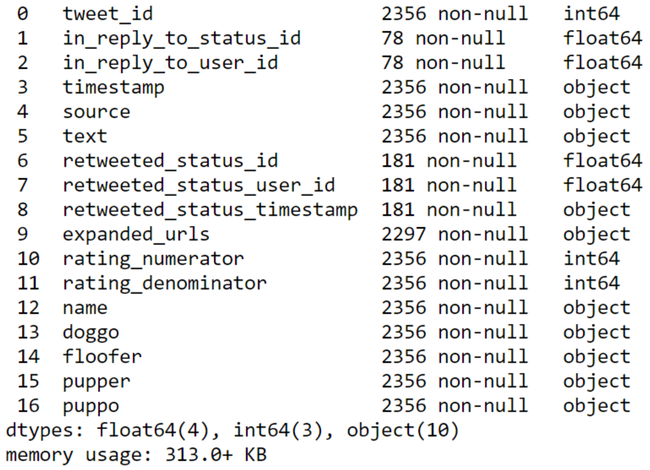
Cleaning Data

This is the final step for data wrangling process. There are three steps which are:

- Define: define and suggestion for the issue
- Code: code to rectify the issue
- Test: to validate the desire result whether it is achieved.

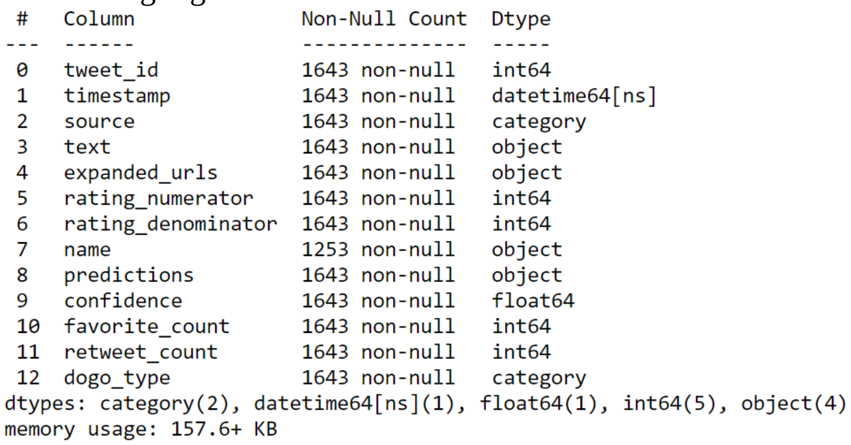
Result

Before wrangling



```
 0   tweet_id                    2356 non-null   int64
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object
 4   source                      2356 non-null   object
 5   text                        2356 non-null   object
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object
 9   expanded_urls               2297 non-null   object
 10  rating_numerator            2356 non-null   int64
 11  rating_denominator          2356 non-null   int64
 12  name                        2356 non-null   object
 13  doggo                       2356 non-null   object
 14  floofer                     2356 non-null   object
 15  pupper                      2356 non-null   object
 16  puppo                       2356 non-null   object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

After wrangling



```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            1643 non-null   int64
 1   timestamp           1643 non-null   datetime64[ns]
 2   source              1643 non-null   category
 3   text                1643 non-null   object
 4   expanded_urls       1643 non-null   object
 5   rating_numerator    1643 non-null   int64
 6   rating_denominator  1643 non-null   int64
 7   name                1253 non-null   object
 8   predictions         1643 non-null   object
 9   confidence          1643 non-null   float64
 10  favorite_count      1643 non-null   int64
 11  retweet_count       1643 non-null   int64
 12  dogo_type           1643 non-null   category
dtypes: category(2), datetime64[ns](1), float64(1), int64(5), object(4)
memory usage: 157.6+ KB
```

The initial dataset provided has contains 2356 row tweets archive. After wrangling, we have 1643 rows and 13 columns of data. The dataset is then storing in a csv file namely ‘twitter_archive_master.csv’ and it is now ready to be analyse.