Project Two Wrangling Report: Twitter Dog Rating Data Investigation

Introduction

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. This report presents the data wrangling steps involved by gathering three datasets: Enhanced Twitter Archive, Tweet Image Predictions, and Additional data using the Twitter API. The additional data was generated as a collection of *json* data using the `tweepy` package. Other python packages includes `pandas` , `requests` , `matplotlib` .

Data Gathering

The retweet and favorite counts was extracted from the collection of *json* data using the tweet_ids (2356) from the Enhanced Twitter Archive as below:

```
add_json_data = [i.split('<|>') for i in open('tweet_json_data.txt', 'r', encoding='UTF-8').read().split('\n') if i != '']
add_json_datalist = []
for i in add_json_data:
    #for tweet ids that may not have any data
    if i[1] == "":
        add_json_datalist.append([i[0], '', ''])
    else:
        idict = ast.literal_eval(i[1])
        add_json_datalist.append([i[0], idict['retweet_count'], idict['favorite_count']])

add_json_df = pd.DataFrame(add_json_datalist, columns=['tweet_id', 'retweet.count', 'favorite.count'])
```

	tweet_id	retweet.count	favorite.count
0	892420643555336193	7018	33838
1	892177421306343426	5303	29353

Data Assessment

Quality and Tidiness issues were identified across the three datasets using

```
twitter_archive_enhanced.columns | tweet_image_data.columns
twitter_archive_enhanced.head()  | tweet_image_data.head()
twitter_archive_enhanced.info()  | tweet_image_data.info()
twitter_archive_enhanced.isnull().sum() | tweet_image_data.isnull().sum()
twitter_archive_enhanced.iloc[0].source
```

The following issues were documented as:

Data Quality Issues

Assessing the data quality across the three datasets

- **1.** `twitter_archive_enhanced` table: *`timestamp`* , *`text`* , *`name`* , *`source`* column names need clarity
- **2.** `twitter_archive_enhanced` table: *`timestamp`* column is a string object datatype
- **3.** `twitter_archive_enhanced` table: *`source`* column contains HTML link residues

- **4.** `twitter_archive_enhanced` table: *`source`* column is a string object datatype
- **5.** `tweet_image_data` table: Missing _ `tweet_id` _
- **6.** `tweet_image_data` table: _ `tweet_id` _ is a string object datatype
- **7.** `tweet_image_data` table: _ `p1_conf` , `p2_conf` , `p3_conf` _ are string object datatype
- **8.** `tweet_image_data` table: _ `p1_conf` , `p2_conf` , `p3_conf` _ columns have values with variable number of decimal places

Data Tidiness Issues

Assessing the data tidiness across the three datasets

- **9.** `twitter_archive_enhanced` table: *`doggo`* , *`floofer`* , *`pupper`* , *`puppo`* columns are dog stages expanded into four columns
- **10.** `twitter_archive_enhanced` table has additional information in the `add_json_df_clean` and `tweet_image_data_clean` tables. The _ `tweet_id` _ is the common reference/column between the three tables

Data Cleaning

Copies of the imported datasets were generated before any of the identified issues were resolved as below:

```
twitter_archive_enhanced_clean = twitter_archive_enhanced.copy()
tweet_image_data_clean = tweet_image_data.copy()
add_json_df_clean = add_json_df.copy()
```

Data Quality Issues

Resolving the data quality across the datasets by

- **1.** Renamed *`timestamp`* column to _ `tweet_timestamp` , `text` *column to* `tweet_text` , `name` *column to* `dog_name` , `source` *column to* `tweet_source` _ in the `twitter_archive_enhanced` table
- **2.** Converted *`timestamp`* column to datetime in the `twitter_archive_enhanced` table
- **3.** Extracted the specific sources of the tweet by removing the HTML residues in the *`source`* column in the `twitter_archive_enhanced` table
- **4.** Converted *`source`* column to categorical data in the `twitter_archive_enhanced` table
- **5.** Dropped Missing _ `tweet_id` _ row in the `tweet_image_data` table (Only 1 row is affected)
- **6.** Converted _ `tweet_id` _ column to integer in the `tweet_image_data` table
- **7.** Converted _ `p1_conf` , `p2_conf` , `p3_conf` _ columns to floats in the `tweet_image_data` table
- **8.** Rounded off the values in _ `p1_conf` , `p2_conf` , `p3_conf` _ columns to 3 decimal places in the `tweet_image_data` table

Data Tidiness Issues

Resolving the data tidiness across the datasets by

- **9.** Merged the *`doggo`* , *`floofer`* , *`pupper`* , *`puppo`* columns into 1 column `dog_stages` in the `twitter_archive_enhanced` table
- **10.** Converted the _ `tweet_id` , `retweet.count` , `favorite.count` _ colummns to integer in the `add_json_df` table; Merged the `add_json_df_clean` and `tweet_image_data_clean` tables unto the `twitter_archive_enhanced_clean` table using the _ `tweet_id` *columns as reference, and Converted the* `retweet.count` , `favorite.count` _ columns to integer datatypes

The data cleaning process followed the Define-Code-Test pattern for each of the documented quality and tidiness issues as below:

2. *`timestamp`* column is a string object datatype

Define

- Convert *timestamp* column to datetime in the `twitter_archive_enhanced` table

Code

```
twitter_archive_enhanced_clean['tweet_timestamp'] =
pd.to_datetime(twitter_archive_enhanced_clean['tweet_timestamp'])
```

Test

```
twitter_archive_enhanced_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 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   tweet_timestamp             2356 non-null   datetime64[ns, UTC]
 4   tweet_source                2356 non-null   object
 5   tweet_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  tweet_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: datetime64[ns, UTC](1), float64(4), int64(3), object(9)
memory usage: 313.0+ KB
```

The final master dataset were saved for analysis and visualization

```
twitter_archive_enhanced_master.to_csv('twitter_archive_master.csv', index=False)
```