We Rate Dogs Wrangling Project

The wrangle project is based on dog ratings and data was sourced from Twitter. For the project, three datasets were used which were subsequently merged to form one dataframe. To execute the project, relevant libraries were imported such as pandas and requests, but to mention a few.

The first dataset was manually downloaded from Udacity, **Enhanced Twitter Archive,** which in the project goes by the term '**archive**' for simplicity's sake. The file was read using the pandas method of pd.read_csv.

The second data set is an **image prediction** dataset that was programmatically downloaded and read as **image_pred**.

The final dataset was downloaded manually from udacity, **tweet-json.txt.** Due to mobile verification issues, the data was not accessed from the Twitter API. The data was loaded and saved as **df_api.**

Assessing Data

Quality issues

1. Columns in wrong datatype

i) Id columns in int or float data type not string;

- tweet_id
- in_reply_to_status_id
- in_reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id
- id

ii) columns meant to be in datetime datatype in object type,

- timestamp
- retweeted_status_timestamp
- created_at

2. Null values misrepresented as "None" across dataframes.

3. column names not descriptive

- jpg_url
- img_num

- p1
- p1_conf
- p1_dog
- p2
- p2_conf
- p2_dog
- p3
- p3_conf
- p3_dog

4. Retweets and replies in dataframes, only need original tweets.

5. source column not human readable.

6. Values in names column (archive dataframe) that are not actual names.

7. Inconsistent case format.

- 'p1'
- 'p2'
- 'p3'

8. Outlier values in rating_numerator and denominator columns

Tidiness issues

1. Repetition of columns

- in_reply_to_status_id : in_reply_to_status_id_str
- in_reply_to_user_id : in_reply_to_user_id_str
- quoted_status_id : quoted_status_id_str
- possibly_sensitive : possibly_sensitive_appealable

>>> repetition across all three dataframes

- Source
- in_reply_to_status_id
- in_reply_to_user_id
- entities and extended_entities are a little more complex with information compounded from other columns

2. No observations in geo,coordinates, and contributors columns(df_api dataframe)

3. doggo, floofer, pupper, puppo columns in archive dataframe need to be merged into one column

1.

Solutions

To ease cleaning, the tidiness issues were addressed before the quality issues.

	Problem	Solution
1	**Repetition of columns**	Dropped all repeated columns
2	**doggo, floofer, pupper, puppo columns in archive dataframe need to be merged into one column**	Merged the 3 columns into one column
3	**source column not human readable**	Used regex to extract portions which were human readable.
4	**Inconsistent case format**	Converted all values to uniform case (title case)
5	**Values in names column that are not actual names**	Used regex to extract those values and replaced them using np.nan.
6	**Retweets and replies in dataframes, only need original tweets.**¶	Looked for values where retweets were null and dropped the rest of the values.
7	**Null values misrepresented as "None" across dataframes.**	Replaced the “None” string using np.nan
8	**Columns in wrong datatype**	Assigned appropriate data types to columns using astype method and pd.to_datetime for datetime columns
9	**column names not descriptive**	Renamed columns with non-descriptive names to descriptive names to

		make it easier to understand the data.
10	**Outlier values in rating_numerator and denominator columns**	Used the IQR to eliminate outliers.

Storing the Data

The resulting data frame was stored as a csv file called “twitter_archive_master.csv”. All further analysis was performed on this file.