Data Wrangling Report

Project objectives

The project main objectives were:

1. Perform data wrangling (gathering, assessing and cleaning) on the provided sources of data.
2. Store, analyze, and visualize the wrangled data.
3. Reporting on

Step 1: Gathering Data:

i. Download (twitter- archive_enhanced.csv') manually.
ii. Gather The tweet image predictions and we will make that programmatically by request with (url).
iii. Gathering data from twitter programmatically by Tweepy

Step 2: assessing Data with cleaning :

Assessing by Quality and tidiness

1:df_arch (name) have none instead of NaN and delet uniuque names
Solve: use drobna and replace unique values

2:df_arch (expanded_urls) has NaN values.
Solve: use drobna to delete because they are not valid data.

3:df_arch (source) Change provided URLs to the corresponding 4 categories.
Solve: there is duplicated values so use replace for specafic values

4:df_arch ('doggo', 'floofer', 'pupper', 'puppo') have none instead of NaN.
Solve: use replace with np.nan to match with dataframe

5:df_arch (timestamp) is str instead of datetime.
Solve: use – pd.to_datetime to convert type(date) to use it for time.

6:df_arch (rating_denominator) has values less than 10 and values more than 10 for ratings.
Solve: use Removed any rows with denominator more than 10

7:df_api (created_at) column is str instead of datetime.
Solve: use – pd.to_datetime to convert type(date) to use it for time

8:df_arch (rating_numerator) make type float and extract int from str.

Solve: Extracted the rating score correctly and converted it to float

9:df_api (id column) name different than the other 2 data sets.

Solve: rename the column

## tidiness

1:df_image (img_num) is not needed.

Solve: remove column by drop column

2:df_api (Just 3 columns needed id, retweet_count, favorite_count)

Solve: Removed other columns

3:df_image (the columns (p1, p1_dog, p1_conf, ...etc)) should be just breed and confidence.

Solve: to make suitable for data frame and useful

4:All datasets should be combined into 1 dataset only.

Solve: Combined all the 3 datasets into one pndas df

Step 3: visualizing :

Make visualize to show data for analyze