Data wrangling Project

1. Gather: I gathered data from 3 different sources and created a DataFrame for each dataset retrieved.

2. Assess: The assessment was carried out with both visual and programmatic inspection.

a. Quality

I. Completeness:

Visual:	"None" values in df1 for puppo, pupper, doggo and floofer columns.
Programmatic:	"None" is considered as a value (string) and not "NaN".

II. Validity:

Programmatic:	Denominator in column rating_denominator should be "10".
Programmatic:	Only leave the rows without retweets from df1 (only original posts).

III. Accuracy:

Programmatic:	Check the Regex for rating_numerator column in df1 (missing floats).
Programmatic	Check dog names with Regex in df1.
Visual:	"Source" column in df3 should only contain the platform information.

IV. Consistency:

Programmatic:	dtype of timestamp column in df1 should be in time format.
Visual:	Standardize dog's name (Uppercase) in p1, p2 and p3 columns of df3.

b. Tidiness:

Visual:	drop "entities" columns from df3 with irrelevant nested information.
Visual:	Transform "doggo, floofer, puppo, pupper" columns into one column (same variable in different columns).
Programmatic:	Only leave the rows without retweets or replies from df1 (only original posts).
Programmatic:	Merge 3 dataframes into 1 master dataframe and save it as a CSV file.

3. Clean: the issues found from the assessment were the following:

- Replace "None" values for "NaN" in columns "puppo", "pupper", "doggo" and "floofer" of df1.

- Correct every denominator value greater than 10 in column "ratings_denominator" from df1.

- Compare the rating_numerator values with my own RegEx value. Replace th column if they are different.

- Change datatypes from object to datetime for "created at" and "timestamp" columns from df3 and df1.

- Standardize breed names in p1, p2 and p3 columns from df2 (uppercase and space instead of underscore).

- Delete every record that is a retweet or reply from df1.

- Apply a new RegEx and correct dog names in df1.

- Extract the platform information the tweet was created from.

- Create one column for dog stages in df1. The columns "puppo", "pupper", "floofer" and "doggo" are related to the same variable.

- Merge the 3 datasets into one Dataframe using inner joins. Before joining, I will discard the columns with no use for this project.