Wrangling and Analyzing Data

Data Gathering

Data for the project came from three sources:

- Original twitter archive data: csv provided via email
- Predictions data: programmatically downloaded from Udacity
- Addition twitter data: obtained from the Twitter API using Tweepy

Inclusion Criteria

Before cleaning, three inclusion criteria were developed to screen the available data:

1. Only tweets with images were included
2. No retweets were included
3. No replies were included

Data Cleaning

Multiple quality and tidiness issues were identified for the three tables. Missing data issues (Issues 6 &

7) were addressed first. Tidiness issues were addressed second and the remaining quality issues were addressed after this. While the quality issues, an additional tidiness (Issue 20) and two additional quality (Issues 21 & 22) were identified. These were addressed as they arose.

Details of the issues identified and solutions developed are found in the following tables.

Quality

1	Retweets are included in the dataset	Remove rows that have reply or retweet information. Remove non-shared “tweet_id” from predictions table.
2	Replies are included in the dataset	Handled in solution for Issue 1.
3	Erroneous datatypes existed in multiple columns, typically IDs not as string, and string not as datetime	Change “tweet_id” to *str* and “timestamp” to *datetime* datatypes.
4	The “exapanded_urls” column contained missing data	Handled through solution for Issue 1.
5	Missing data in the columns “name” to “puppo” were classified as the string “None” not NaN	Handled in solution for Issue 6.
6	Missing counts for the columns “doggo” to “puppo”	Use *for* loop and *.str.contains()* to identify if text contains each column header. Include text if it is found. If not, return NaN.
7	Missing names for the “names” column	Create function to identify pet names and re-populate “name” column.
8	Some entries in the “names” column were not names	Handled in solution for Issue 7.
9	The “text” column contained a shortened hyperlink as well as the tweet text	Create a function to remove links and apply it to “text” column.
10	A second name existed for some tweets and was not identified	Was not handled.
11	The “rating_numerator” column contained incorrect values	Create a function that identifies the value before the last / in “text” column. Apply this to “rating_numerator” column. Manually correct any ratings not covered.
12	The “rating_denominator” column contained incorrect values	Create a function that identifies the value after the last / in “text” column. Apply this to “rating_denominator” column.
13	Erroneous datatype for “tweet_id” column	Change “tweet_id” to *str* datatype.
14	The reduced count of entries for this table indicates that some entries in the archive do not have images	Remove any tweet ids in the archive table that aren't in the predictions table.
15	Erroneous datatype for “tweet_id” column	Change “tweet_id” to *str* datatype.
16	Retweet and favorite data is missing for some tweets and cannot be retrieved	Handled through solution for Issue 1.

Tidiness

17	Multiple columns contained the same type of data, e.g. “doggo” to “puppo” all contain dog type info	Create a column “dog_type” and fill with column data in order of puppo, pupper, floofer, doggo using .fillna(). Drop the redundant columns.
18	Multiple columns contained the same type of data, e.g. “p1” to “p3” all contain dog breed predictions	Change column names to “prediction_#”, “confidence_#”, and “dog_#”. Use *pd.wide_to_long* to collapse each type into a single column.
19	The data in the api data table should have been connected to the other tweet information	Use a left join to merge api data with archive data on “tweet_id”.

Results

The final cleaning resulted in two tables.

Archive Table

1971 observations across 12 columns

Predictions Table

5913 observations (3 predictions per tweet) across 7 columns