Data Wrangling: WeRateDogs™ Twitter Account

Data Gathering

Data was gathered from three sources:

1) I read a .csv file of tweet data into Jupyter Notebook using pd.read_csv
2) Image prediction data was downloaded programmatically from Udacity
3) I queried the Twitter API using OAuth to get data on retweet count, favorite count, and tweet length.

Data Assessing

I assessed the data both visually and programmatically, collecting issues in a list organized by table and issue type (tidiness or cleanliness).

Data Cleaning

The issues identified and addressed are listed and described in the table below. In general, I took care of missing data, then tidiness, and finally cleanliness issues.

Table	#	Issue	Fix	Type
Archive	1	78 replies/retweets included in dataset	Used df.drop to drop tweets that were replies or retweets.	Cleanliness
	2	Tweet text contains url	Wrote function using str.split() to separate text from URL. Put URL in new column.	
	3	Erroneous data types	Dropped unnecessary columns, then employed .to_datetime() and .astype() where needed	
	4	Missing names & incorrect name recognitions Some tweets are about two dogs second name of dog missing (not addressing)	Function written using regex to accept the tweet text and return correct name. Success was not 100%, but better.	
	5	Missing classifications for doggo, floofer, pupper, or puppo	Used loop to identify text in tweet and append classification to list, then place in new series	
	6	Presence of more than one dog category in tweet	Addressed in 5	
	7	Series for "We only rate dogs" reprimand unavailable	Wrote function to recognize tweets with the words “We Only Rate Dogs,” then created boolean column to store results.	
	8	doggo, floofer, pupper, puppo violate "Multiple columns shouldn't contain the same type of data"	Created new column for dog stage in 5, dropped four redundant columns.	Tidiness

Image Predictions	9	Many predictions are lowercase and contain underscores	Used short function to remove underscores and capitalize all predictions.	Cleanliness
	10	There are fewer image_preds entries than there are tweets in the archive (perhaps a result of retweets, perhaps not)	Examined overlap of image and archive dataframes, eliminated all tweets not present in both.	
	11	Erroneous data type (tweet_id is integer, not string)	Addressed in 3	
	12	p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog, p3_dog violate "Multiple columns shouldn't contain the same type of data"	Melted various columns into single columns	Tidiness
API Data	13	Erroneous data type (tweet_id is integer, not string)	Addressed in 3	Cleanliness
	14	There are 17 fewer entries in api_data than there are in archive	Resolved when api_data is merged with archive	
	15	This dataset is separate from the rest of the data	Merged API table with archive using a common key.	Tidiness