Wrangle Report

(Derrick Xu, 07 June 2020)

This write-up documents the steps taken to wrangle the WeRateDogs datasets. Data wrangling is a three-step process: Gathering, Assessing and Cleaning.

Gathering

In the first step, three separate datasets were gathered from three different sources. The table below summarize the datasets, and the methodologies used to extract the data:

	Name	Description	Methodology
1	WeRateDogs twitter archive	.csv file containing basic tweet data for all 5000+ of their tweets	Imported using pandas pd.read_csv function
2	tweet image predictions	.tsv file containing dog breed prediction for every image in the WeRateDogs Twitter	Imported using pandas pd.read_csv function
3	WeRateDogs Tweet JSON	Additional data in JSON obtained via Twitter's API	Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Write each tweet's JSON data into a .txt file. Read line by line into a pandas DataFrame

Assessing & Cleaning

The second step is Assessing. In this step, two attributes of the data are identified: quality and tidiness. Using both visual and programmatic assessment, eight quality issues and three tidiness issues were identified.

It is recommended that quality issues related to completeness be tackled first, before tidiness and then rest of the quality issues. As such, the WeRateDogs issues were tackled in this sequence: Tidiness issues before quality issues. This is because thankfully, the data gathered is relatively complete. For example, no missing data observed for the columns of interest in WeRateDogs twitter archive (i.e. tweet_id, rating_numerator, rating_denominator).

The quality and tidiness issues are summarized in the tables below. Have included the third step Cleaning as well for easier reference. The final cleaned dataframe is saved as a .csv file and used for drawing insights and data visualization.

Quality Issues

	Dataset	Assessing	Cleaning
1	WeRateDogs Twitter Archive	Many columns are not useful for analysis (e.g. in_reply_to_status_id, in_reply_to_user_id, timestamp, source)	Drop columns that are not useful, and keep only meaningful ones (tweet_id, rating_numerator, rating_denominator)
2		Only interested in original tweets, but there are retweets	Remove the retweets by checking for retweeted_status_id
3		Inaccurate numerators & denominators due to the presence of more than one fractions in tweet	Correct inaccurate numerator and denominators manually
4		There are ratings given with decimal place (eg. 13.5/10)	Will correct numerators & denominators manually
5		Other than point #3, there are more denominators not equal to 10. Will need to remove outliers	Drop tweets with denominator != 10
6	WeRateDogs Tweet JSON	Many columns are not useful for analysis	Keep only meaningful columns, especially favorite_count, retweet_count
7		Only interested in original tweets, but there are retweets	Remove the retweets by checking for retweeted_status_id
8	Tweet Image Prediction	Many columns are not useful for analysis	Keep only meaningful columns
9		There are 66 duplicated dog pictures urls	Will remove duplicates
10		To analyze dog breed, we need to condense the data	Will condense image prediction into 1 column, based on px_dog and px_conf, using df.apply()

Tidiness Issues

	Dataset	Assessing	Cleaning
1	WeRateDogs Twitter Archive	Columns on dog stages (doggo, floofer, pupper and puppo) do not fulfill the tidiness criteria "each variable forms a column"	Will merge the dog stages into a single column called "dog_stage"
2	WeRateDogs Tweet JSON & Tweet Image Prediction	The 3 datasets are part of the same observational unit. Will need to fulfill the tidiness criteria "each type of observational unit forms a table"	Will create a master dataset "df_all_clean" by merging all 3 datasets together