The data wrangling process:
gather, assess, and clean.

Maha Alossiami
22 June 2020

Introduction:

The dataset used in this project is the tweet archive of the WeRateDogs Twitter account. WeRateDogs rates dogs. The goals of the project are to wrangle and analyze the collected data.
Python 3 was used including these libraries:

- pandas
- NumPy
- requests
- tweepy
- json

Gathering data :

The data was collected from three different sources, using a provided CSV file, an URL, and Twitter data.

Assessing data

After gathering the data, some errors discovered using visual assessment and programmatic assessment. These errors are fatal and needed to be fixed.

- **Quality issues**

Including missing values, duplicated, and invalid data.

The tweet archive data

- Retweets are included in the data and must be removed because we are only interested in tweets with ratings.
- columns with NaN values are unnecessary
- Invalid name as a and none are false data
- Expanded_urls missing values

- source column incorrect format and hard to read

The tweet image predictions data

- 66 duplicated data in the jpg_url column.
- p1, p2, and p3 contain inconsistent names format.

The tweet data

- To make a consistent form the id column renamed to be to twitter_id as the rest of the datasets.

- **Tidy issues**

Tidy issues are related to the data structure

The tweet archive data

- The timestamp contains a date and time.
- Four columns for dog type (doggo, floofer, pupper, and puppo) instead of one.
- join the three data sets into a master data set

Cleaning data

Each problem stated above were fixed using the appropriate data manipulation tools. In the case of including retweets in the data set, rows were deleted. Columns that contain NaN values that cannot be replaced were removed. The formate of columns changed if necessary. Combining columns was used to create one column with concise information. Lastly, to create the master data set, I join the three sets by index.