Wrangle and Analyze data

Introduction

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent."

Key points

Key points to keep in mind when data wrangling for this project:
We only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
Cleaning includes merging individual pieces of data according to the rules of tidy data.
The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.

Part 1
Project Details

Data wrangling, which consists of:

Gathering data
Assessing data
Cleaning data

Assessing Data

Once the three tables were obtained I assessed the data as following:

- Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.

- Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

Quality issues

which includes Completeness, Validity, Accuracy, Consistency :

- unusual names for dogs like None,a,bo etc
- numerator and denominator in ratings that are not according to rules
- datatype for timestamp column in archive dataset
- columns in archive like retweeted_status_id ,retweeted_status_user_id etc.
- datatype for datetime column in tweet
- user_favourites,user_followers are redundant columns in tweets dataset
- missing values as number of rows not equal in all datasets

Tidiness issues

which includes structural issues :

- stage variable in four columns: doggo, floofer, pupper, puppo
- three different datasets for same data 'df_tweet' and 'df_image' and ‘df_archive'

Cleaning

Cleaning our data is the third step in data wrangling. It is where we will fix the quality and tidiness issues that we identified in the assess step.

First and very helpful step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original.

Whenever I made a mistake, I could create another copy of the dataframes and continue working on the cleaning part.

How to tackle Quality issues

which includes Completeness, Validity, Accuracy, Consistency :

- remove retweeted_status_id where not null
- mine numerator and denominator ratings that are not according to rules
- convert datatype for timestamp column in archive dataset
- remove columns in archive like retweeted_status_id ,retweeted_status_user_id etc.
- convert datatype for datetime column in tweet
- remove user_favourites,user_followers in tweets dataset
- fill in or remove wherever necessary missing values as number of rows not equal in all datasets

How to tackle Tidiness issues

which includes structural issues :

- convert stage variable in four columns: doggo, floofer, pupper, puppo into one columns by melting.
- Merge 'df_tweet' and 'df_image' to 'df_archive' to facilitate cleaning