WeRateDogs Twitter Analysis Project

Data Wrangling Report

The project's motivation was to create interesting and trustworthy analyses using WeRateDogs data.

# About the Data

The data collection procedure was divided into three stages:

* All I had to do was download the csv file after receiving the twitter archive data.
* The image prediction data was extracted from udacity's server using the requests library.
* Utilizing Tweepy library, tweet data containing favorite counts and retweet counts was extracted as a json file using the tweet id available in the archived file.

Pandas was used to import all datasets into the working environment.

Before cleaning, the data shape was (2356, 17), (2075, 12), and (2326, 4) respectively. The first figure represents the number of rows, and the second figure represents the number of columns.

# Quality Issues

The following quality issues were discovered during the assessment phase:

The majority of the dog names recorded in the archive tweet data were incorrect. There were 745 dog names that contain None instead of nan, and the rest were split between correct names and some stop words. There was some redundancy, as 78 of the records were replies to other users, and 181 were retweets of the original tweets. Some columns had incorrect data types, such as the timestamp and tweet id columns. Outliers were found in the ratings denominator and numerator columns. Not all images are for dogs according to the p1 predictions in the image prediction data and tweet\_id had the wrong datatype across all dataset.

To address the quality issues, the incorrect data types were corrected, as this was critical for the merging stage. The column mean was used to populate outliers, none recorded was replaced with nan, and redundant features were removed.

# Tidiness Issues

There were no obvious tidiness issues.

For dog type, multiple columns were used in the archive data. For example, 'doggo,' 'floofer,' 'pupper,' and 'puppo.' This was fixed by combining the columns into one.

Even when there were no null values, there was some inconsistency in the number of unique tweets: 2075 and number of unique image url: 2009. This only confirms that retweets were present in the data.

After carefully fixing all observed issues, the 3 datasets were merged into one in readiness for the analysis phase. The data shape after cleaning and merging was (1341, 15)