Udacity: WeRateDogs Project Analysis

Introduction

WeRateDogs is a twitter handle that rates dog pictures that people post on twitter. In this report I will be describing the data wrangling steps that I took to gather and clean the data for this project.

Data Gathering

There was a csv file that was provided by Udacity that had the tweet-ids for the reviews made by WeRateDog's twitter handle. There was also an additional tsv file that was programmatically downloaded from Udacity's server.

Apart from this to fetch the retweet count and the favorite count, Tweepy API was used. The returned object was a JSON object which was stored locally as a text file.

All of this data was read into a dataframe using Pandas library.

Data Assessing and Cleaning

The data assessing and cleaning step was done together for each issue that was found out. There were 2 types of problems that were found in the data-

A. **Tidiness Issue:**
   1. **Table Merge:** The JSON content file that was downloaded from twitter was merged with the csv file to form a single dataframe that has all the tweet information.
   2. **Redundant columns:** There were 4 columns (doggo, pupper, puppo and floofer) that had information about the category of the dog. These 4 columns were merged into a single column that had this information. Some dogs that were in both doggo and puppo category had a new category ('dogg+ other') assigned to them.

B. **Quality Issues:**
   1. **Timestamp-** It was found that the timestamp and retweeted_status_timestamp column were in non-null string format. They were changed to datetime format for easier analysis later on.
   2. **Retweet data-** Retweets should be excluded from the analysis as they don't have the original ratings. These rows were deleted.
   3. **Denominator column-** There was some discrepancy between the denominator column data and the ratings from the tweet. This discrepancy was mostly due to the regex search finding a different pattern in the tweet text that wasn't the rating. These tweets were cleaned.
   4. **Numerator column-** There was some discrepancy between the numerator column data and the ratings from the tweet. These discrepancies were either due to the regex pattern search error or due to decimal ratings. These ratings were also corrected.
   5. **Rating Column-** Since the denominator for the ratings were different, a rating column was required that changed all the ratings to a base of 10. Made a new column consisting of (rating_numerator/ rating_denominator) *10
   6. **Missing rating-** Some rows did not have any ratings in them and were general tweets. These had been excluded.
   7. **Modifying source column-** The source column was modified so as to remove the hyper link.

8. **Category Datatype-** The source and category columns should be ‘category’ datatypes rather than strings. These were modified.
9. **Extra category due to uppercase-** In the dog breed neural network table one of the dog breed, ‘Cardigan’, was duplicated and was also written as ‘cardigan’. This was corrected.

After all this, the dog breed table and the tweet tables were combined to form a master dataset. We did this for easier analysis later on. This master table was also exported as a csv file.