Wrangle-Report

1. Wrangling Report

Prepared by Akash Yadav

1.1 Gathering

In the gathering step, we are asked to gather 3 different tables using 3 different ways: 1. twitter-archive-enhanced.csv: this file is obtained by downloading it manually from the udacity site. 2. image-predictions.tsv: this file is obtained by downloading it programmatically using the python's requests library 3. tweet_json.txt: this file is obtained by calling the twitter API for each tweet, using the tweet id as parameter, which we obtained from the twitter-archive-enhanced.csv data. Instead of using the Requests library, we use Tweepy library which provides a higher level way to get the data from the API. After that, we make sure that all the data are available in the same directory as the wrangle-act.ipynb

Tidiness

All retweets and replies should be deleted The image predictions could be condensed to show just the most confident dog breed prediction. All three data frames can be combined into one single dataframe.

1.2 Assessing And Cleaning

After gathering all the 3 files, their data is stored into a data frame for easier assessment and cleaning. In order to assess the data, I examined it visually and programmatically using python's pandas library. First, we printed out all the data frames entirely, used the info() function to assess the datatypes, used describe() function to summarize the quantitative variables in the datasets, etc. Then examined the data frames more specifically by examining each variable separately and found out the following issues.

1)Corrected the wrong data types (timestamp, retweeted_status_timestamp) • Removed records that was a retweet or a reply • Merged Doggo, Floofer, Pupper and Puppo columns into one column to look like this.

2) I removed some unnecessary columns that i am not going to use • Then I removed 23 records that have denominator Not equal to 10

3) I classified my numerators into 3 types o Normal (10 to 15) since most of the ratings have numerators in this range o Low (Less than 10) I found some ratings less than 10, but I needed to know if this is a typo or this is a real rating o
Outlier (More than 15) some of the records have very high numbers more than 1000.

4)I compared a random sample of my low numerators to the text and I found that it's low on purpose, I decided to keep it as part of the analysis. For the Outlier High numbers, I decided to remove then because even if they are high on purpose, they will distort my analysis and skew some averages.
Then I changed the missing dog names from None to NaN
• Last but not least, I merged the 3 datasets together.

5) The cleaned data frame was saved to a new CSV file.