Weratedogs Twitter Archive - Wrangle Report

Athier Aljahdali

Data Gathering

I gathered data from 3 sources, stored in separate files:

1. Weratedogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, I downloaded by using the file in Udacity

Library.

The favourite_count and retweet_count was extracted programmatically from this file. I loaded the 3 raw data files into separate tables: archive, predictions and json_data.

Assessment & Cleaning

I tried to Assessment data using .head() , shape function , find out the duplicated() and snull().sum()

Cleaning

1)Twitter archive data

Quality:
1 There are 23 cases where the denominator of rating != 10. These entries will be removed.
2 remove many cases of where the rating_numerator is smaler than 10
3 remove many cases of where the rating_denominator is smaler than 10
4 Rename columns with more appropriate names: "timestamp" to "tweet_timestamp", "text" to "tweet_text", "rating_numerator" to "dog_rating_out_of_ten", "name" to "dog_name"
5 Since retweets and replies will be removed, the column "retweeted_status_timestamp" will be removed as it will no longer provide any useful information.

6 Remove column "rating_denominator" once all the values that != 10 have been removed since this will no longer provide any useful information.

Tidiness:

1 There are 181 retweets which need to be removed. All columns related to “retweets” will be removed, we can drop all columns related to retweets.
2 No need to all the coulmn in images dataset just tweet_id and jpg_url what we need
3 There are 4 kinds of source the users used We can replce the it with the type
4) drop rating_denominator and rename rating_numerator to rating
Iphone
Vine
Twitter Web
Tweetdeck
5) The data type of timestamp should be datetime (string)

2)df_tweet

Quality&tidiness :
1 After trying to merge the data, it appears that there is some non-numeric values for the "tweet_id" inputs which will need to be removed.

3) image_predictions

Quality & tidiness :

1 Missing values from images dataset (2075 rows instead of 2356)
2 Some tweet_ids have the same jpg_url