Wrangle Report - Project: WeRateDogs

Introduction

This document describes the data wrangling effort on the project – WeRateDogs. The objective was to take in the given dataset, clean and tidy it up so that one can do a realistic analysis to get some actionable insights

The Data wrangling effort consisted of following stages:

1. Gathering data
2. Assessing data
3. Cleaning data

Gathering

The Data for this Project was gathered from three sources as described below:

1. **The enhanced WeRateDogs Twitter archive file**
   This file was manually downloaded from by clicking twitter_archive_enhanced.csv

2. **Tweet data from Twitter using Twitter API**
   Using the tweet IDs in the WeRateDogs Twitter archive, the Twitter API was used to query Twitter server and download each tweet's JSON data.

   A Tweepy library function was used to fetch and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

3. **Data on the prediction of the breed of Dog breed**
   This file (image_predictions.tsv) contains 3 predictions of the breeds of dog (or other object, animal, etc.) from the dog images associated with each tweet. This prediction has been done by a neural network. The file image_predictions.tsv is hosted on Udacity's servers and was downloaded programmatically using python Requests library.

All the data have been read in to individual Pandas Dataframes - dftwitt_arch, dftweet_data and dfimage_pred.

Wrangle Report - Project: WeRateDogs

Assessing

After gathering data from each of the above sources, a visual and programmatic assessment was done to identify at least 8 quality issues and 2 tidiness issues. Given below is a list of the Quality and Tidiness issues that were identified and worked upon.

Quality

Data from extended Twitter archive file

1. Retweets are not needed since the images/ratings are repeated - 181 retweets
2. Float Data type for Columns : in_reply_to_status_id,in_reply_to_user_id
3. String Data type for timestamp and retweeted_status_timestamp
4. The ‘rating_numerator’ column has few values unusually low or above 10
5. The ‘rating_denominator’ column has a few values above 10 and below 10
6. 1976 rows do not have any value for dog stages (doggo,floofer, pupper and puppo)
7. 14 rows have more than one value for dog stages instead of 1 - e.g. : doggo and puppo or doggo and pupper)
8. The expanded_urls column has 59 blank values - so no pictures of dog

Relevant Tweet data pulled using twitter API

9. The data has only 2345 tweet Id's compared to 2356 tweet Id's in "twitter archive"

(Note: For Analysis - only columns id, favorite_count and retweet_count will be used)

Data read from Image Prediction File

10. This data has only 2075 tweet ids compared to 2345 and 2356 tweet ids as above
11. 66 jpg image URLs are duplicated

Tidiness

1. In "twitter archive" Multiple Columns for Dog Stages (doggo, floofer, pupper and puppo) are not required – we can use only 1 column
2. Dataframes - dftwitt_arch and dftweet_data should be combined to form one single dataframe.
3. In dfimage_pred - reduce the set of columns for 3 predictions to 1 - Have just a single prediction that would be a dog or ‘None’

Cleaning

All the Quality and Tidiness issues identified in the Assessment stage were verified and fixed to the extent possible.

The cleaning process basically comprised 3 steps:

1. Define – Defined the cleaning action needed
2. Code - Write the code to fix the issues programmatically
3. Test - Verify if the issue was fixed by viewing the output of test code

Before cleaning, copies of the dataframes were made and these copies were operated upon for purpose of cleaning.

After all the cleaning the resultant cleaned data was stored in two files:

1. twitter_archive_master.csv : Stores twitter data
2. dog_image_pred.csv : Stores Dog breed prediction data

Conclusion

The total effort of Data wrangling and Analysis was about 10 days – spent about 3 to 4 hrs/day.

Though the report may show all the issues listed sequentially – they were not identified sequentially and in the assessment stage, some of them were identified during the cleaning process. It was an iterative effort.

As the data familiarity sets on visualizing it through multiple angles, you start noticing more inconsistencies. Knowledge of the domain – helps to do things faster as you are comfortable with the data.

Finally as the insights started showing up – it was extremely gratifying.