Wrangle Report

Introduction

Data from the real world seldom comes clean. The purpose of this project is to wrangle Twitter data from WeRateDogs to create interesting and accurate analyses and visualizations.

The needed effort in this project for my data wrangling consists of:

1. Data Gathering.
2. Data Assessing.
3. Data Cleaning.

1) Data Gathering:

Collecting data from three different sources:

- **Enhanced WeRateDogs Twitter Archive**

Download file twitter_archive_enhanced.csv manually.

For all 5000 + of their tweets, the WeRateDogs Twitter archive contains simple tweet details, but not all.

The archive contains one column: the text of each tweet, which I used to extract ratings, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to "enhance" this Twitter archive.

- **Image Predictions File**

This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

One more cool thing: I ran every picture through a neural network that can identify breeds of dogs * in the WeRateDogs Twitter archive. Results: a table full of image predictions (only the top three) alongside each tweet ID, image URL, and the image number that matched the most accurate prediction (numbered 1 to 4 since up to four images can be used for tweets).

- **Additional Data via the Twitter API**

The suggested step for acquiring this information should be to query the Twitter API for the JSON data of each tweet using the Tweepy library of Python and store the entire JSON data collection of each tweet in a file named tweet_json.txt.

I can't set up a Twitter developer account using the suggested measures, however, so I will collect this piece of data without a Twitter account for this segment.

2) Data Assessing:

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues.

Detect and report at least eight (8) quality problems and two (2) tidiness problems.

❖ Observations of Assessment

A. Quality issues (Completeness, validity, accuracy, consistency)

✓ **twitter_archive**

1) wrong datatype for some columns e.g.: timestamp should be datetime instead of objec (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers/strings instead of float.)

2) There are 181 retweet entries. We only want Use only original ratings tweets, not retweets.

3) Rating denominator and numerators value are inconsistent or incorrect.

(However, rating numerators that are greater than the denominators do not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.)

4) Wrong numerators and denominators are captured from text

some are actually referring to date or other meaning 15/8 and 24/7.

wrongly captured due to decimal and spaces "11. 26/10" was captured as "26/10"

5) Tweets that has no image.

✓ **image_prediction**

6) Inconsistent capitalization on predicted dog names

7) There are 324 rows non dog image (where p1, p2, and p3 are false)

8) Duplicated jpg URLs with different tweet ids

9) 'None' string should be replaced with 'NaN'

10) All 3 files contain different number of rows.

B. Tidiness Issues

1) 2 columns storing rating information.

2) 4 columns (doggo, floofer, pupper, puppo) to indicate dog stages.

3) All 3 files contain common tweet_id column, which can be used to join all three files as one dataframe.

3) Data Cleaning:

Fixing the quality and tidiness issues that I identified recently

The dataframes are copied to another new dataframes before the cleanup begins.

There are 3 phases are involved in the cleaning process:

- ✓ Define: transform evaluations into specified tasks for cleaning. Such concepts also act as an instruction list so that others can look at the work and repeat it (in the future).
- ✓ Code: convert to code those definitions and run the code.
- ✓ Test: Visually or with code, test dataset to ensure cleaning operations have performed.

The steps of evaluating and cleaning data are replicated many times in this project. -- time the steps are repeated, it makes the data more relevant to examine later on.

4) The End:

Wrangled data is stored in the twitter_archive_master.csv. Then, to draw valuable knowledge from it, it is ready to be analyzed, then to create visualization and reports.