Wrangle Report

Ahmed M. Hasan

Our goal in this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Wrangling Data:

This process consists of three steps:

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

1.Gathering Data

We have collected data from three sources:

- The WeRateDogs Twitter archive. The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which was used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, The ratings probably aren't all correct. Same goes for the dog names and probably dog stages (see below for more information on these) too. We'll need to assess and clean these columns if we want to use them for analysis and visualization.
- The tweet image predictions: The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). 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
- Twitter API: Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. Because we have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? We're going to query Twitter's API to gather this valuable data.

2. Assessing Data

We assess the collected data to find out two different issues:

A. Quality Issues
B. Tidiness Issues

A. Quality Issues

The four main data quality dimensions are:

- Completeness: missing data
- Validity: data making sense
- Accuracy: inaccurate data
- Consistency: standardization

B. Tidiness Issues

Three requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

We have done the assessment of data by two methods:

I. Visual Quality Assessment
II. Programmatic Quality Assessment

3. Cleaning Data

A. Quality Issues:
    - Delete retweets in `twitter_dogs`
    - Convert `timestamp` from `string` to `datatime`
    - Convert the lowercase and invalid names to `None`
    - Delete row with the `rating_denominator` of an invalid value
    - Convert `rating_numerator` to float
    - Correct the inaccurate `rating_numberator` by extracting the correct fractional rating
    - Change `tweet_id` from an integer to a string
    - Remove the html tags from `source`, only keep the content inside the tags
    - Remove underscore from `breed_pred`, make all title case
B. Tidiness Issues
    - Merge the clean versions of archive, predictions, and counts dataframes
    - Create one column for the various dog stages: doggo, floofer, pupper, puppo
    - Extract dog breed and prediction confidence into two columns
    - Drop unnecessary columns

For details and code, check the ***wrangle_act.ipynb*** notebook.