Udacity Data Analyst Nanodegree

Project “Wrangle and Analyze Data”/ Description of Data Wrangling

Richard Steele

Gathering the Data

This phase consisted of several steps to ensure collection of the required data. I loaded the necessary libraries into the Jupyter notebook. I used the provided archived file and Udacity url (wherein the image predictions resided; for ease in processing I converted it to a csv format) to create separate dataframes (**rating_archive** and **image_preds**) for each file. After procuring the proper authorization credentials from the Twitter Developer page after creating a developer account, I created an API object for extracting the tweets from the Twitter JSON file and converting said data into a text file **tweet_jason.txt**. The code for this step included parameters for recognizing and waiting for the Twitter extraction rate limits, as well as a list of the tweet-ids for the failed extractions. Finally, I converted **tweet_jason.txt** to a pandas dataframe, **df_tweets**.

Assessing the Data

Most of the assessment phase was visual, via inspection of the csv files from which **rating_archive** and **image_preds** arose. I noticed several columns in both files that contained data which were either missing (e.g., “None” for the maturity-stage designations in **rating_archive**) or irrelevant to my planned analyses (e.g., the descriptive text in **rating_archive**). I assessed all the dataframes programmatically, confirming the visually based conclusions, and looking for any missed nulls and duplicates. In accordance with the requirement to exclude retweets, I created a list by vectorially searching for retweets; the list was used to drop those retweet rows separately from other cleaning steps as it was a presentational prerequisite.

The issues I chose to document are:

Quality Issues

1. image_preds: tweet_id in exponential float format
2. image_preds: inconsistent formatting for columns 'p1', 'p2', 'p3'
3. rating_archive: rating denominator is always 10, so column is redundant
4. rating_archive: missing data in columns 'in_reply_to_status_id' and in_reply_user_id'
5. rating archive: missing data in columns 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp'
6. rating_archive: column 'timestamp' is in 'object (string)' format
7. rating_archive: column 'expanded urls' irrelevant to analysis
8. image_preds: 66 duplicated rows
9. image_preds: lack of consistent and defined schema in dog-breed format

Tidiness Issues

1. rating_archive: dog "stages" ("doggo," etc.) are in separate columns
2. rating_archive: column 'text' contains both rating and "stage"

Cleaning the Data

The cleaning phase for quality consisted mainly in removing rows and columns which were duplicated, irrelevant, and redundant. Changes in formatting were accomplished via vectorization instead of looping. After each cleaning step a programmatic test confirmed that the changes went through properly.