WeRateDogs Twitter Project
Wrangling Report

Uma Maheswari Raju

Introduction:

This section describes about the wrangling procedure followed for WeRateDog Project. The wrangling process includes following steps:

(1) Gathering data
(2) Assessing data
(3) Cleaning data

Gathering Data:

There are three different type of dataset used in this project.

(1) **Twitter_archive_enhanced.csv** – The WeRateDogs Twitter archive. We manually downloaded this file.

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

(3) **tweet_json.txt** – Each tweet's retweet count, favorite and any additional data we found interesting. Using the tweet IDs in the WERateDogs twitter archive, we could query the twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called **tweet_json.txt** file. Each tweet's JSON data stored in a line

Gathering Data : Summary

It is the first step in the wrangling process. **twitter_archive_enhanced.csv** file was read from csv file using pandas. **Image_predictions.tsv** was downloaded using requests function. **tweet_json.txt** was obtained querying API and getting JSON object of all the tweet_ids using Tweepy. After that all data was imported into our programming environment (Jupyter Notebook)

Assessing Data:

After gathering all data from different sources, each type of data was assessed visually and programmatically for quality and tidiness issues. This process includes checking data type, value counts, number of non-null entries and duplicates. Quality issues are assessed considering completeness, validity, accuracy and consistency of data. Tidiness issues are assessed by considering the following the Hadley Wickham guidelines:

1. Each variable must have its own column
2. Each observation must have its own row
3. Each value must have its own cell
4. Each "kind" of variable in a data set(linked with the others)

After assessing through visual and programmatic assessment, the following quality and tidiness issues were found

Quality issues:

Archive dataset:

- misssing data in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp, expanded_urls
- timestamp, retweeted_status_timestamp are object instead of datetime
- some dog names are invalid names such as 'None', 'a' a, an, the, just, one, very, quite, not, actually, mad, space, infuriating, all, officially, 0, old, life, unacceptable, my, incredibly, by, his, such
- few rating_numerator and rating_denominator numbers are very high numbers. Also tweet_id such as 786709082849828864 has numerator value as erroneous one which is different from numerator value in 'text'.
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be string instead of float
- tweet_id is int instead of string

Images dataset:

- tweet_id is int instead of string

json_tweets dataset:

- 14 tweet_ids are duplicates.
- date_time is object instead of date time
- tweet_ids are int instead of string

Tidiness

- archive dataset has too many dog breed columns. instead it could be one column with dog stage in archive dataset
- tweet_id in other two tables need to be removed when we combine all tables

Cleaning Data:

After assessing quality and tidiness issues for all datasets, the next stage is cleaning data. In this stage we fix all quality and tidiness issues that we identified in the assessing step. Before cleaning, the all datasets were copied. The process was define, code and test.

Archive dataset:

1. misssing data in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp, expanded_urls
   ⇒ for some tweet ids, above variables are NaN. In this project, the available values for the above variables are considered and NaN values are kept as it is.
2. timestamp, retweeted_status_timestamp are object instead of datetime.
   ⇒ Following variables were changed to date time using pd.to_datetime function
3. tweet_id is int instead of string
   ⇒ changed using .astype
4. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be string instead of float
   ⇒ above variables changed to string using .astype
5. some dog names are invalid names such as 'None', 'a' a, an, the, just, one, very, quite, not, actually, mad, space, infuriating, all, officially, 0, old, life, unacceptable, my, incredibly, by, his, such
   ⇒ changed to 'None'
6. few rating_numerator and rating_denominator numbers are very high numbers. Also tweet_id such as 786709082849828864 has numerator value as erroneous one which is different from numerator value in 'text'.
   ⇒ Added new columns 'extracted_numerator' and 'extracted_denominator' in archive_clean data frame after regex based match of extracted numerator and denominator values from 'text' column.

images dataset:

7. tweet_id is int instead of string
   ⇒ changed using .astype

json_tweets dataset:

8. 14 tweet_ids in json_tweets are duplicates
   ⇒ Removed using drop duplicates

Tidiness

- archive dataset has too many dog breed columns. instead it could be one column with dog stage in archive dataset
    - ⇒ breed columns changed as one breed column using melt function
- tweet_id in other two tables need to be removed when we combine all tables
    - ⇒ tables are merged into one dataset using merge