WeRateDogs Twitter Data
Wrangle report
Manal Almehmadi

In the project, I follow Data wrangling process Gathering, Assessing, and Cleaning

Gathering Stage

I gathered data from three different sources. The first source is `twitter-archive-enhanced.csv`. It had the major chunk of the data about tweets of the WeRateDogs account from 2015 to 2017. Second source is a file that was to be programmatically downloaded from the Udacity servers which had the results of the machine learning algorithm 'neural network' performed on the images from the WeRateDogs account.
I downloaded this file using Python library **requests**. The third source for gathering data was `web scrapping off Twitter` using Tweepy API. The Tweepy API is an easy to use Python-based API which connects to a twitter account using secret and public keys then read json file and extract the required fields from each tweet's JSON data and store it in a separate file, **tweet_data_extra.csv**, for use during the assessment stage. I had issues with Tweepy API, the file that scraped tweets off twitter is empty even after implement the code, so I follow the second approach which is read JSON file (tweet_json.txt) that provided by Udacity.

Assessing Stage

In this stage, I did both programmatic and visual assessment for the data. Programmatic assessment using python functions (info, describe, value_counts, head and duplicated). I used these functions to

- info( )- To check data types of columns and null entries.
- describe( ) - To numeric summaries
- Head( ) - To show top row
- duplicated( ) - To find duplicated rows based on all columns

I discovered main eight quality issues and two tidiness issue.

Cleaning Stage

First, I made a copy from the original data frame, then start solving tidiness issues

1. I combined all three tables (**tweets_df , img_predictions and archive_df tables)** because they are related all of them about tweets.

2. I merged the dog stages **(doggo, floofer, floofer, and puppo columns)** into a single column.

3. Correct data type of below columns
    - tweet_id from int to Object
    - in_reply_to_status_id and in_reply_to_user_id from float to object
    - dog_stage from object to categorical type
    - Timestamp from object to datetime
    - Retweets and favorites from float to int

4. Fill null in Retweets and Favorites columns with median values.

5. Replace incorrect dogs name (a, an, the, just, one, very, quite, not, actually, mad, space, infuriating, all, officially, 0, old, life, unacceptable, my, incredibly, by, his, such) with correct one from available text.

6. Use regular expression to extract the source of tweet from JSON text in source column.

7. Drop tweets with rating_denominator NOT equal to 10

8. Extract the correct rating from text column useing regular expression
9. Fix rating values for some tweets; according the available text
10. Remove not original tweets which is retweeted tweets

For each issue, I identified the issue and the way I am trying to solve it, start coding, then test it to make sure the issue is solved.

Storage

I Stored the clean data frame in a CSV file named twitter_archive_master.csv