Udacity data analysis nanodegree

Data Wrangling Report

The main goal of this project is to achieve a complete data analysis process (gather, access, and clean) of twitter WeRateDog account. In addition, to analyzing this data and provide some valuable answers from it.

- ## Gathering data:

  1- Twitter archive and predication image which were provided from Udacity.
  2- For Json file it was supposed to be downloaded programmatically but I had an issue with twitter to provide me a developer access. For previous mentioned reason Json file which was provided by Udacity was used.

- ## Accessing data:

  The three data frames were checked individually for any issues (quality, accuracy, or tidiness issues). From these issues we were requested to mention 8 issues and later fix them in the cleaning step.

  The 8 issues I worked on were:

  For twitter archive data there were 6 issues:

  1- There are a lot are 181 NaN values in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp, which mean there are missing data (1810) in these columns.
  2- tweet_id supposed to be object not integer (as mentioned by reviwer)
  3- By checking data info, expanded urls show 59 missing values.
  4- For (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id) data type supposed to be int64 rather than float64.
  5- For timestamp retweeted_status_timestamp have wrong data type it supposed to be datetime rather than object.
  6- By checking name values it shows that there are some mistakes such as a, an, the.
  7- Rating supposed to be at maximum 10. It can be seen in previous code that there is some incorrect rating.

  For image data:

  8- It shows that there are only 2075 values rather than 2356. That means these columns have 281 missing values.

  For Json file:

  9- CrateDate has wrong data type (object rather than datetime).

Done by: Ohoud Almadani

Tidiness issues:

1- Three data frames could be joined in one for easier analysis.
2- Four dogs' stages could be joined under one column (as suggested by reviewer)

- **Cleaning:**
  - To solve issue 1,3 unnecessary columns (in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp','expanded_urls') were dropped.
  - To solve issue 2, (after merging all data frames) tweet_id data type changed from integer to object as suggested by reviwer.
  - To solve issue 4, timestamp data type changed to datetime.
  - To solve issue 5, incorrect names I used regex (python regular expression) as suggested by reviewer.
  - To solve issue 6, incorrect rating, a new column was generated using formula of (10*numerator/denominator) and then a definition of maximum 10 was apply on it. Then datatype changed to integer to remove decimal points.
  - To solve issue 7, archive and image data were merged based on ID to drop any missing rows.
  - To solve issue 8, CrateDate data type was changed to datetime.
  - To solve tidiness issue 2, four dog's stages ([['doggo', 'floofer', 'pupper', 'puppo') merged in one column as suggested by reviewer.
  - To solve tidiness issue 1, all three data set were merged in one data set (twitter_archive_master) and saved as csv file.

Udacity data analysis nanodegree

Done by: Ohoud Almadani