Wrangle Report

About this project :

This project (WeRateDogs) shows the data wrangling process for the tweet archive of Twitter user @dog_rates (aka WeRateDogs). This user has over 4-million followers and has received international media coverage. In this project, data wrangling technieques are used to gather, assess and clean the twitter archive for this user.

Gather Data:

Three different files are used in this project to analyze the data.

Enhanced Twitter Archive :

This file has the basic twitter data (about 2500 tweets) about the tweets and contains the info about dog 'stages' and other information and this is provided in a .csv file named **twitter-archive-enhanced.csv**

Image Predictions File:

This file has probability of image predictions for each tweet id and also has image url. Date for this file is available in **image_prediction.tsv**

Retweet and Favorite count File:

Data for this file collected by querying Twitter API for the above user and stored in **tweet_json.txt** file

Assess Data:

All three files were assessed for data quality and tidiness using Pandas' dataframe package. Following are the findings of quality and tidiness issues:

twitter archive data

Quality Issues

- Several columns for this data have missing info
  1. in_reply_to_status_id (clean them)
  2. in_reply_to_user_id (clean them)
  3. retweeted_status_id (clean them)
  4. retweeted_status_user_id (clean them)
  5. expanded_urls (clean them)
- several columns have 'None' value (clean them)
- Some of the values in 'name' column start with lower-case and also contain inappropriate values for 'name' (such as : 'a', 'an', 'quite' etc.) (clean)

Tidiness

- dog types (i.e., doggo, floofer, puppoer and puppo) can be be combined into one column (tidiness)
- Some of the rows have classified the dogs into two types (e.g., same row has value both in doggo and floofer columns) - (tidiness)
- some of the rows in expanded_url have reference to different urls separated by commas (tidiness)
- 'text' column has a combo of description , ratings and shortened urls (tidiness)

image prediction data

- some of the names in predictions table are lower-case (clean them)

Clean up of the data

Quality and tidiness are handled using pandas library by :

1) Dropping noisy columns from the files
2) Removing any duplicate rows
3) Combining three dataframes (corresponding to each file) into a single dataframe