Data Wrangling

Gather

Here we have gathered data from multiple sources and sourced it from three type of files.

CSV

“Twitter-archive-enhanced.csv”, this file was readily available as it was downloaded from Udacity. It was then sourced into panda df.

TSV

“Image_predictions.tsv” file was sourced into df by downloading it on local machine using response package, which helped us extract it via link programmatically rather than manually downloading it.

TXT

“Tweet_json.txt” was extracted using twitter API then saved in program, then saved in local machine, then extracted back.

Asses

Objective

Before assessing I had defined the objective so that clean would be done accordingly.

1 - What is the highest rating given
2 - Value vs count of the rating, mode of rating here, along with visualizing the same
3 - Which tweet received the maximum number of likes and what was it's rating

While extracting data only handful of data, approx 50 was extracted using twitter API, as it would have taken 33 mins to extract the complete set. However keeping the objective in mind, it wouldn’t matter.

Visual

This was going through the data, noticing things like names had none values, retweets were present, lot of unnecessary columns were present too.

Programmatic

Here null values were identified and dropped, values less than 10 for numerator_rating were dropped as no dog is less than 10. Removing false predictions in the p_dog column.

Quality parameters that were detected

Null values
Ratings less than 10
Retweets
Names with value None
Favorites with value 0
Image number column as data type, float
Incorrect dog breed
Predictions with False values in p_dog

Tidiness

Drop all irrelevant columns
Merged all the tables in one for convenience

Clean

Above assessed parameters were then worked on to obtain a clean data.

Quality

Dropped all rows with null values
Dropped ratings less than 10
Retweets texts were dropped as well
Name values with None was dropped.
Favorites with value 0 was dropped.
Data type of img_num column was changed to integer from float.

Tidiness

All the DF were merged in single named final.
It had the following columns only as the rest were dropped:
tweet_id
Text
rating_numerator
rating_denominator
favorites name
retweets
jpg_url
img_num
p1
p1_dog

Limitations

We should keep in mind the process that a dataset of only 50 entries was worked on. This result may not be appropriate for making generalized statements but it fulfills our purpose.