WRANGLE & ANALYZE DATA

Iintroduction:

in this project, we are going to gather data about the account in twitter known as WeRateDogs its user @dogs_rates, this account rating dogs using people comments by uploading photos of the dogs and let the people rating it.

What we did in this project:

- Data wrangling, which consists of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analysing, and visualizing

Gathering data:

We gathered the data from these three sources:

1. The WeRateDogs Twitter archive. we Download this file (twitter_archive_enhanced.csv) was provided by Udacity to students

2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers, we downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. And this file tweet-json.txt, we should use Twitter API and Tweepy library, but we could not get the consumer and secret keys from twitter developer.

Assessing data:

In these files, we find quality issues and Tidiness issues it declared below：

Quality

twitter_arch:

- We are not going to use these columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id and retweeted_status_user_id) so,we will drop this columns

- Retweeted_status_timestamp, timestamp should be datetime instead of object

- The numerator and denominator have unacceptable values.

- name column have unusual names

- Source column unreadable

- delete expanded_urls we are not going use URL in analysis

image

- Missing values in images its 2075 rows instead of 2356 as in twitter_arch

- tweet_id should be string

Tidiness:

- Dog type variable in four columns: doggo, floofer, pupper, puppo

- merge to dataframes (image),(twitter_arch) and (tweet_data)

Cleaning data

In this step we solve and fixing the issues that we find it in the assessing step programmatically using library's, also we following this iterate (Define, code and test) in the first we copy the data frames and then using this iterate to clean the data

Storing, analysing, and visualizing

After finishing the clean step and completely no issue is there then we stored the data in a file named twitter_archive_master to commit our work there.

Then we ask sample questions for analysis and visualize.