Wrangle Report

1. Introduction

This project is about dealing with some data for famous account on twitter called WeRateDogs the user publishes the images for doges and share it on twitter and make rate so we are clean the archive and extract data for those dogs tweets and also open images dog file and clean it also use API tweeter to gather retweet count and favorite count as well after collect those data we need proceed wrangling process to clean the data that we have and then do some analyzing and investigation on the data.

2. Gather Data

- In first dataset [twitter-archive-enhanced.csv] I just import data to jupyter notebook and read it by b abanda library
- For 2nd dataset it was online file on the internet so I use requests library and download the file then open it and later read it by panda
- For 3rd dataset I create tweeter developer account to gather the data use API then save it as Jason file then convert it to csv file and read it.

3. Assess Data

In this section I try to explore each dataset ether visualize or Programmatically for looking to data that need to fix and clean it later.

Assessment Issues

Quality

- Twitter Archive (Tweet_Arch data)
  1. comparing Tweet_Arch data with Imges_pred data 2356 compare to 2075 it same there is missing rows I believe this tweet without images.
  2. I notice in name column there is value like (a, an, the, such, o,etc ..) is not really names is seem like extracting issue
  3. in_reply_to_status_id, in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp some of them. contain NAN value which is not useful data.
  4. There are some rows have more than one dog.
  5. name variable have some values None it must be change to NAN.
  6. Missing data in column(expanded_urls).
  7. Erroneous datatypes tweet_id , timestamp,source,rating_denominator.

- Tweet image predictions (Imges_pred data)
  1. some of jpg_url is duplicated.
  2. tweet_id convert to str.
  3. drop this columns p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'.

- json API (df_tweets data)
  1. tweet_id convert to str.

Tidiness

1. Dog types column is separated in three columns.
2. we need to join and combined all datasets in one dataset

4.conclusion

After we do cleaning I store dataset in new file for visualize and analyze the data , before I cleaning I find missing rows comparing with dataset of image so I try to join the datasets together , there is duplicated image I eliminate them from image dataset also there is unstructured data like doges type it was separated into three columns so I use code to set it in one column beside a lot of missing data for some columns I just drop them also there is incorrect values I try to extract text column to gather the correct names.

Finally, I believe we can iterate those process [gather- assets – cleaning] again until we satisfy on the result.