Data Wrangling

1- Gather

Data were collected from three different sources. First data was collected from the “twitter-archive-enhanced.csv” file which was in the same directory in which project notebook was located. The csv file was imported into pandas dataframe.

Second data was extracted programmatically *The URL*, Python’s request library was used to extract data from *URL*. *The URL* was split using “/” as the separator and the last value was the file name. This file was written in the content of our request. Then this file was imported as a dataframe in pandas using tab as the separator. The dataframe was named “images_pred”.

The third data, actually I couldn’t do this step appropriately, so I just downloaded the file from the classroom, but I know how to do it, the data should be extracted from Twitter API using python’s tweepy library. You need to extract the favorites and retweet counts for each tweet. This data should be saved as a JSON file using UTF-8 encoding.

The images dataframe, the JSON file and the archive data were should be into a single data frame, A copy of this merged data was saved in CSV format.

2- Assess

In this section I try to explore each dataset ether visualize or Programmatically looking to data that need to fix and clean it later using multiple methods like: .head() , .info() , . value_counts() …etc.

3- Clean

After exploring the dataframes I have come with serval issues:

Quality

- Several columns have empty values:
  - *in_reply_to_status_id*
  - *in_reply_to_user_id*
  - *retweeted_status_id*
  - *retweeted_status_user_id*
  - *retweeted_status_timestamp*.
  - *date_time*
- There are some rows contains more than one dog.
- The timestamp column is an object. It has to be a datetime object.
- There's some of url images are duplicated in *images_pred*
- tweet_id convert to string
- unequal rows between images dataframe and archive dataframe, whitch means that row doesn't include images.
- In several columns, null values are not treated as null values.

Tidiness:

- Dog types column is seperated in three columns.
- we need to join and combined all datasets in one dataset.