Dataset

A twitter archive from We rate dogs is taken as the data set the date set contains inherent problems as it is taken via web scraping through a twitter API and other resources, Since data is inconsistent and pose many issues such as redundant columns wrong data types etc hence , cleaning is required before the data set could be used for Analysis .

Issues

- ## Tidiness

1. Many columns like doggo , floofer ,pupper , puppo essentialy desc the same thing(dog stage) and should be in a single var named dog_attribute .

There are redundant columns such as doggo , floofer ,pupper , puppo which depicts the same thing ie. Dog stage and hence a new var is created as dog_attribute which capture dog stage details. Later dog_attribute is changed to a categorical data type.

2. Join tweet_info , image_pridiction to tweet_archive.

A join operation is used to combine all the above tables.

3.Get rid of retweeted_status_id retweeted_status_user_id retweeted_status_timestamp.

Retweets are not to be analysed and hence are removed via drop.

Wrangling Report

- Quality

1. Get rid of rows with expanded urls as NaN.

Tweet is useless without a picture and hence Urls with NaN are droped.

2. Dogs names are inconsistant(first letter not caps or wrongly named).

There are issues with dog names such as first letter is not in caps , some names are wrongly assigned . to clean this the issues are gathered in a list and rows with wrongly assigned names are located and changed (using reg ex) via for loop .

3. Display full contents of the text column.

Text column is the back bone of the dataset ie most of the columns are derived from Text basically text is the tweet itself . due to pandas default settings text is truncated and to display full text using pandas set_option method

4. Dog names in case not there should be NaN not None.

Dog names are assigned as none in 2nd issue . to resolve this simply replace 'None' with NaN

5.Data contain retweets(will take care in tidiness last part).

Retweets are not needed and hence removed

6.source rendered as html tags.

Source is in the form of full Url enclosed within html tags which are inconvenient to read and hence using replace will do the trick

7.ratings must have decimal value where required.

Some rating contains decimal value but rating_numerator does not have a decimal value to correct this can be cleaned via same method as 2nd issue.

8.fix incorrect ratings .

When there are two '#' / '#' formats in text then wrong ratings might be taken up hence using reg ex and for loop will clean this issue and give right ratings

9.make rating numerator as float.

Data type of rating reated columns should be float

10.Errorneous Data types (timestamp , tweet_id , in_reply_to_status_id , in_reply_to_user_id).

Correct the wrongly assigned data types.