Wrangling, Assessing and Cleaning Efforts

Gathering data

In this section , we will gather data from different resources like :

** CSV files*

In this type of data we gathered The “twitter_archive_enhanced.csv” file that was provided to me in udacity lessons

** TSV files*

In this type of data we gathered The “image-predictions.tsv” file will is downloaded programmatically using the requests library and the given URL

** APIs*

Using Python's Tweepy library to access Twitter API

The steps to access twitter api

The steps to query data from twitter api are as follow:

1- Get a developer account from twitter.

2- Use the account credentials to set an access token to open a connection with the api.

3- Query the api to get the data for the tweets which we have their IDs in the archive file.

4- The data is in json format, we store this data in a txt file.

5- We open the file and get the needed information from it using the key name.

Assessing the data

We assessed the data both visually and programmatically to determine its issues which needs to be cleaned

Visual assessment includes using some functions as dataframe.head () – dataframe.tail ()

Programmatic assessment includes using some functions as dataframe.info () - dataframe.describe () – dataframe.series.value_counts () - dataframe.is_duplicated () ...etc.

The issues we want to determine are classified into Quality issues and Tidiness issues

The issues in the fathered data are :

Quality Issues :

- There are some retweeted tweets and we only care with original tweets
- Time stamp is string and not a date time format
- The denominator has some values other than 10 although it must be a constant number equals 10
- The rating numerator is of type int and should be converted to float because some ratings have floating point
- The numerator has some values less than 10 which contradicts with the rule of the account rating that they "they are good dogs Brent." , and some other values are very large that seems to be not reasonable for example 1776
- some names of the dogs are missing and written 'None' aand some maybe Typos and written 'a' or some other words that begin with small letters like 'one' , 'officially'..etc.
- There are duplicated image URLs corresponding to different tweets id
- some columns won't be used in analysis and can be dropped after cleaning duplicates like
- Some columns we will not use them in our analysis like 'in_reply_to_status_id,in_reply_to_user_id,source

Tidiness Issues :

- There are 4 columns at the end which represent the stage of the dog (doggo,floofer,pupper,puppo) these columns represent the same variable and must be converted to the one column only' , and the tweet which has multiple stages will be converted to string “multiple”
- The data frames can be merged with each other as they form the same observational unit

Cleaning data

After the data assessment, We took a copy of all data frames so that the initial data frames are kept with us , and then we started to clean the data , and un each cleaning step we use following methods:

Define the issue

Here we are defining a clear statement for the issue, and providing a method to solve it

Clean the issue

Here we are using python and pandas functions to query the data, delete duplicates, delete non found or biased data or group data together

Test the issue

Here we are testing that the issue that we defined earlier is totally solved