We Rate Dogs Twitter Data Analysis

There are three steps in Wrangling of data
i)Gathering of Data
ii)Assessing of Data
iii)Cleaning of Data
I will give brief intro about these three

i)Gathering of Data:-Gathering is a process of collecting the data.Data can be collected from various source like from database ,website,excel file,csv.
Data can be download manually or programatically.When you have thousand data to gather from various sources.It is difficult to do manually as well as this is time consuming.So It's better use program to download data.
ii)Assessing the Data:-Assessing is the process of detective agent.We have to found out quality issue and tidiness issue in our data.
Quality:-It is content issue.It could be like missing data,proper name space.
Tidiness:-It is structural issue.Some irrelevant information is present.

Gathering of Data

We have gathered Data from 3 sets.
i)twiiter archive inhanced data has been download manually.It is hosted in Udacity server
ii)image prediction fill we have downloaded data programmatically using request library
iii)twitter json data has been dowloaded using tweepy api.

```
import tweepy

consumer_key = 'YOUR CONSUMER KEY'
consumer_secret = 'YOUR CONSUMER SECRET'
access_token = 'YOUR ACCESS TOKEN'
access_secret = 'YOUR ACCESS SECRET'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)
```

Assessing the Data

I have tried to go through the data manually and programatically like a detective to found out the issue in them.

I have several issue in main Data ‘ twitter-archive-enhanced.csv’ and there is also some issue in Image-predictions.tsv
I will mention some quality issue of twitter-archive-enhanced.csv file then I will mention some quality issue of ‘Image-predictions.tsv’

These are quality issue of twitter-archive-enhanced.csv file

Quality

Twitter Archive table

1.Remove (+0000)from timestamp
2.Remove (+0000)from retweeted_status_timestamp
3.Convert timestamp datatype to DateTime object
4.Convert retweeted_status_timestamp datatype to dateTime object
5.In text remove the incorrect http link
6.Some of Rating numerator are very high
7.Name of dog is not same as in tweet.
8.Remove underscore and do propercase of name of dogs

Image Prediction

9). proper case and remove underscore from dogs name

Cleaning

We have mainly used pandas library to clean our data.

Twitter Archive table

1.Remove (+0000)from timestamp:-
We have used string replace method for replacing ‘+0000’ from timestamp column

2.Remove (+0000)from retweeted_status_timestamp
We have used string replace method for replacing ‘+0000’ from retweeted_status_timestamp column

3.Convert timestamp datatype to DateTime object
We have used pandas.to_datetime object to convert string object to datetime object of timestamp column

4.Convert retweeted_status_timestamp datatype to dateTime object
We have used pandas.to_datetime object to convert string object to datetime object of retweeted_status_timestamp column

5.In text remove the incorrect http link

We have tried to find out incorrect in tweet text .We have find the lenght of the tweet.If the length of http link is less than 15 then it must be incorrect link .So we have removed incorrect http links.

6.Some of Rating numerator are very high
We have removed outliers of rating numerator .Since some of the rating are very high like 1500,200,1600 etc..
We have removed all the ratings above 13.Since most of the users have rated their dogs less than equal to 13 out of 10.
We have used function to iterate through each row find the rating .If rating is above 13 drop the row.

7.Name of dog is not same as in tweet.
We have fetch name of dod from tweet text by using key word like ‘This is|Meet ‘ .In all most of all the cases after this name of dog comes.We tried to match with existing dog name in name column .If it does not match .We will replace it with tweet dog name

Image Prediction

9). proper case and remove underscore from dogs name
We have properly aligned naming convention.We have removed underscore and make capital letter each word.

Tidiness

Twitter Archive

1)Merger four columns doggo,floofer,pupper and pupoo to make single column called species.Since all reprsent certain kind of Species

Image Prediction

2) drop image num (there is no significance of column)