Wrangle-and-Analyze-Data Project

Prepare By : nouf alfayez

1/ Gathering Data

In this project, I worked on the following three datasets:

1- **Enhanced Twitter Archive**

I downloaded this file from Udacity source, then I worked to import this file as dataframe (df_t).

2- **tweet image prediction**

I downloaded this file from Udacity source, then I worked to import this file as dataframe (df_img).

3- **Data from Twitter API**

I didn't use twitter API to import data,because I don't want to create the developer account, just Itake this data from Udacity source, then I worked to import this file as dataframe (df_json).

2/ Assessing Data

Visual Assessment:

I used this code lines ‘For three dataframes’to discover any quality and tidiness issues:

```
df_t
```

In [7]:

Out[7]:

	source	timestamp	in_reply_to_user_id	in_reply_to_status_id	tweet_id	
	a> com/download/iphone" ...r	2017-08-01 16:23:56 0000+	NaN	NaN	892420643555336193	0
	a> com/download/iphone" ...r	2017-08-01 00:17:27 0000+	NaN	NaN	891777421306343426	1

```
df_img
```

In [14]:

Out[14]:

	p1	img_num	jpg_url	tweet_id	
	ger_spaniel	1	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	666020888022790149	0
	redbone	1	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	666029285002620928	1

```
df_json
```

In [19]:

Out[19]:

	url	retweet_status	retweet_count	favorite_count	tweet_id	
	https://t.co/MgUWQ76dJU	Original tweet	8853	39467	892420643555336193	0
	https://t.co/0Xxu71qeIV	Original tweet	6514	33819	891777421306343426	1

Quality issues:

1. 'df_t': I will remove these columns (puppo, pupper, floofer, doggo) and change it with one column with name(type).
2. 'df_t': The (timestamp) column Recorded in a different format.
3. 'df_t': Not necessary the html tags in (source) column.
4. 'df_t': Unify the values of (rating_denominator) to be equal 10.
5. 'df_t, df_img, df_json': Convert (tweet_id) column from integer to String
6. 'df_img': Convert (img_num) column from integer to String.
7. 'df_img': Rename the columns (p1, p1_conf, p1_dog
   , p2, p2_conf, p2_dog
   , p3, p3_conf, p3_dog) to be more clearly.
8. 'df_img': Remove duplicates in jpg_url column

Tidiness issues:

1. I will remove these Columns in dataframe(df_t):
   in_reply_to_status_id
   in_reply_to_user_id
   retweeted_status_id
   retweeted_status_user_id
   retweeted_status_timestamp
   because it have alot of null vaues

2. Merge all data frames (df_t, df_img, df_json) into a single data frame named (df).

3/ Cleaning Data

In this step of the project, I fixed all quality and Tidiness issues, first I made copies of each of the three dataframes. I named the cloned dataframs with these names (df_t2, df_img2, df_json2), and then I first started working on cleaning up the problems of Tidiness:

1/ I combined all the dataframes under a single datafram with this name (df).

2/ I deleted some columns with too many of missing data.

After that I started working on cleaning uu the quality issues.

4/ Storing Data

After I finished cleaning datafram (df), I stored it in a csv file with the name twitter_archive_master.csv