Datasets:

twitter-archive-enhanced.csv:

This dataset is archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. *(1)

image-predictions.tsv:

The tweet image predictions are present in each tweet according to a neural network. *(2)

tweet-json.txt:

This is the resulting data from twitter_api.py. *(3)

Sources: *(1)

https://classroom.udacity.com/nanodegrees/nd002-mena-connect/parts/71de6fde-0474-4933-85c8312aa416cbfe/modules/74066a17-93c0-4033-8638-8019c456dc3a/lessons/e31b008a-4fac-4591-8f2a-4a5c8dddf445/concepts/5e3db54a-1a5f-41a6-8e20-fd99f201861d

(2)

https://classroom.udacity.com/nanodegrees/nd002-mena-connect/parts/71de6fde-0474-4933-85c8312aa416cbfe/modules/74066a17-93c0-4033-8638-8019c456dc3a/lessons/e31b008a-4fac-4591-8f2a-4a5c8dddf445/concepts/5919f3b1-899f-4295-80f1-17f091eb4df6

(3)

https://classroom.udacity.com/nanodegrees/nd002-mena-connect/parts/71de6fde-0474-493385c8312aa416cbfe/modules/74066a17-93c0-4033-8638-8019c456dc3a/lessons/e31b008a-4fac-45918f2a4a5c8dddf445/concepts/d7e3de1b-d7a1-4ebc-9d58-beba021a7c29

Python Libraries:

I used several libraries in this project to a gathering, assessing, visualsation data, request, cleaning data…etc.

They are as follows:

- pandas.
- requests.
- json.
- re.
- numpy.
- Image.
- requests.
- BytesIO.
- matplotlib.pyplot.
- seaborn.

Data wrangling process:

1- Gathering Data:

twitter-archive-enhanced.csv:

I downloaded this dataset from udacity classroom and saved it as pandas data frame to manipulate with it.

image-predictions.tsv:

I requested this dataset from an url by using request library and saved it as a variable then I wrote it in tsv, and in the last, I read this notepad as pandas data frame separated by tab to manipulate with it.

tweet-json.txt:

I have some issue with a tweeter when I send an email they are late to transfer my account in a tweeter to a developer to make me use a tweeter API and because I am late to submit my project I download tweet-json.txt from udacity to complete submitting this project locally.

I loaded this dataset as a json in the variable, because I don't need all of the tweet-json dataset variables -columns- in this analysis, and chose the next columns:
id : this is the primary key to make some relationship with other datasets in this analysis.
favorites: total of the preferences of the specific tweet. retweets:
total of the retweets for the specific tweet.
To read this dataset by used open built in, and I chose the important columns for this analysis.
Then I get the specific columns, and save the data in a new list

2- Assessing data:

I assessed data by two ways:

- Visually:

  By using Microsoft Excel.

- Programmatically:

  By using Python libraries and functions like pandas, numpy, and function(info(), sample(), .value_counts()….etc)

Quality:

twitter-archive-enhanced:

- data type of 'timestamp', and 'retweeted_status_timestamp' column is object -string- he must be change to datetime.

- 'in_reply_to_status_id','in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', and 'expanded_urls' these columns has a null values.

- 'in_reply_to_status_id','in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', float and it must be string.

- There are some duplicates in a column 'expanded_urls', and he is not important column.

- all the 'in_reply_to_status_id','in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' must be dropping because it is not necessary in my analysis.

- 'source' column must be transformed from HTML format to the simple string.

- timestamp must be split to date and time.

- 'expanded_urls' has a null value I will drop all rows that hava null value.

- splitting the uncorrected data from the dogs_stage like doggopupper, doggofloofer, and doggopuppo by comma.

image-predictions:

- 281 records are missing.
- some values have '_' instead of space.

tweet-json:

- 2 recordes are missing.

Tidiness:

- 'doggo', 'floofer', 'pupper', and 'puppo' unnecessary I think I Can merge in one column.

- merging the three datasets into one dataset and deleting unnecessary columns.

3- Cleaning Data:

twitter-archive-enhanced:

- transforming 'source' column from HTML format to the simple string and extract the tag content.

- data type of 'timestamp' column is object -string- he must be change to datetime.

- 'in_reply_to_status_id' and 'in_reply_to_user_id' must be converted from float to string.

-Because the following columns have NaN value:

- dropping the following columns because I don't need to them in my analysis, and it has a null values -missing values- 'in_reply_to_status_id','in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', and 'expanded_urls'.

- timestamp must be split to date and time.

- dropping timestamp.

- There are some duplicates in a column 'expanded_urls', and he is not important column.

- 'expanded_urls' has a null value I will drop all rows that hava null value.

- splitting the uncorrected data from the dogs_stage like doggopupper, doggofloofer, and doggopuppo by comma.

image-predictions:

- replace _ with space in p1,p2,p3 columns

- rename the following columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, and p3_dog.