Project: Wrangling and Analyze Data

1 - Data Gathering

Used 3 sets of data for analysis:

a - twitter_archive_enhance.csv, which was downloaded directly;

b- image_prediction.tsv accessed with the help of request.get(url);

c - tweet_json via Udacity website and tweet_extract.

2 - Assessing Data

Dataset data checked: columns, lines, datatypes, info, head, empty lines, errors in general.

Quality issues

'twitter_archive' Table - Checked 7 adjustments needed for the dataset:

1 - Erroneus datatype in 'timestamp' column and remove the characters '+0000';

2 - Retweets and replies should be removed;

3 - Dogs' name with 'None' and other characters;

4 - Unnecessary hyperlinks;

5 - Necessary to change the datatype column tweet_id;

'image_prediction' Table

6. Convert text to lowercase letters in column types of dogs (p1, p2 and p3);

7. Drop duplicates jpg_url (images).

'tweet_json' Table;

8 - Change name column;

9 - Tweets without images.

Tidiness issues

1 - Move twitter_json_clean and image_prediction_clean to twitter_archive;

2 - Substitute columns for one column;

3 - Cleaning Date

1 - Remove '+0000 and convert str/object to datetype in 'timestamp' column;

2 - Remove retweets and replies and respective columns;

3 - Remove 'None' and other characters;

4 - Remove hyperlinks (https://stackoverflow.com/questions/13682044remove-unwanted-parts-from-strings-in-a-column?nodirect=');

5 - Change datatype from;int64 to object;

6 - Convert text to lowercase;

7 - Drop duplicates jpg_url (image);

8 - Changed name of column id_str in tweet_json_clean to tweet_id;

9 – Delete tweets without images;

Tidiness

1- Substitute columns 'doggo', 'floofer', 'pupper' and 'puppo' for one column;

2 - Merge columns tweet_id (tweet_json_clean/image_prediction_clean) to twitter_archive_clean.

4 - Storing Data

Save gathered, evaluated and cleaned master dataset to csv file “twitter_archive_master.csv";

5 - Analyzing and Visualizing Data

Produced three insights and three views.

Insights:

1- The most used means to access twitter is the Iphone, responsible for more than 98% of accesses;

2 - Charlie, Lucy and Cooper are the most popular names;

3 - Have peak retweets around school holidays and Christmas and New Year holidays.