DATA GATHERING
DATA ASSESSING
DATA CLEANING
DATA STORAGE
DATA VISUALIZATION

INTRODUCTION
In this project we are going to be working with WeRateDog twitter account, this rate people dogs It has a constant denominator(10) although the numerator is sometimes larger than the denominator. In this project we are going to be working with an archive file with about 2356 rows, image prediction file with about 2075 rows and additional data that contain both favorite count and the retweet count. This file contains 2354 rows.
This archive contains basic tweet data (tweet ID, timestamp, text, etc.)

AIM
The aim of the project is to analyzing and visualizing dogs that are not retweeted

In this report I will not talk about data visualization, but I will give detail explanation on data Wrangling, which comprises of Data gathering, assessing, cleaning and storage,
Data Gathering
I was provided with the twitter achieve data in a comma separated value file this contain 2356 rows with 16 columns which includes(tweet_id,timestamp,source,text,name,in reply columns,retweeted columns,rating_numerator,rating_denominator,dog stages(doggo,floffer,pupper,puppo)

I used a request function to download the image prediction file to my local machine and read the file as csv to pandas dataframe. This file contains 2075 with 11 columns which is (tweet_id,jpg_url,image_num,breed columns(p1,p2,p3) algorithm confidence column(p1_conf,p2_conf,p3_conf) and prediction that its a breed of dog column(p1_dog,p2_dog,p3_dog)

The twitter counts file, I was unable to get my developer account approved by twitter so i used the alternative method provided by udacity. In this case I also used the requests function to download the .txt file into my local machine and I used pandas.read_json to read the json text file into pandas dataframe. The data set also have 2354 with 30 columns
For the purpose of this project I will only need id,is_quote status,favorite count retweet count, favorited, retweeted columns

Data Assessment

1. The in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns all appear to contain many null (NaN) values. If a value is present in any of these columns, it indicates that the tweet is actually a reply (the first two) or a retweet (the last three). Because one of our requirements

is that "we only want original ratings (no retweets)", we should remove any rows that have non-null values in any of these columns. Aftward, we can drop these columns

2. a href tag at the source column, is not actually needed, we only needed the media to which the tweet was sent, i.e. from an iphone, the website or via another app?
3. The rating numerator column recorded 0 as rating value this is not a problem though
4. At the name column some dogs as weird name like a, an, etc
5. at the dog name, dog stage the null values are written as None which pandas read it to be non null values
6. The dog breed column i.e p1,p2,p3,has inconsistent naming of breed, some are written in title format
7. At the jpg url column there is duplicated image with different tweet_id
8. The doggo column, floffer,pupper,puppo columns suppose to be in a single column named dog stage
9. The p1,p2,p3,and their respective algorithm suppose to be in two columns which are dog breed, algorithm confidence
10. So many columns with wrong data types

Data Cleaning

Before commencing the cleaning process, make a copy of each data frame, we are going to use the copied data for cleaning

1.I selected the columns needed for my tweeter_count data frame because i don't actually need the whole 30 columns, so i only need (id,is_quote_status,favorite _count retweet _count, favorited, retweeted) columns

2. I merge twitter_archive table with my twitter_count table, using a left merge and drop the id column

3. Remove all the non null values at the in_reply columns and retweeted_status column, this will eliminate the retweet and solve the 0 values in favorite and retweet countess column

4.Drop the 2 in_reply, 3 retweeted_status,columns also drop the is_quote_status, retweeted and favorited column because they are no more needed and having them in the table will make the table choking

5. I used python function and apply method to extract the tweet source from the html tag at the source column

6. Extracted the rating numerators with decimal values, using regrex function and extract function, locating them with the .contains function in the text columns. I also extract only text from the initial text column eliminating url and ratings from the column

7. Using find method in excel spreed sheet, some rating are not still extracted so i used manual method to correct that

8. Drop null rows at the text column

9.Drop rows that are not rated dog using match function.

10. Replace the None at the doggo,floofer,pupper,puppo Columns with ‘ ’, Using melt method then create a dog stage column that contain all the four columns,Replace the ‘ ’ at the doggo,floofer,pupper,puppo Columns with NAN,drop the doggo,floofer,pupper,puppo columns

11. Remove the null at the expanded url

12. Remove the img_num col from the image_pre table

13. breed naming inconsistent, some are lower case, some are names in upper case, some are title format. So i convert all the entries in the breed columns to lowercase

14. merge the p1,p2,p3 as breed, p1_conf,p2_conf,p3_conf as pred_confidence this obey the rule of tidiness, one variable forms a column

15. Merge the clean image prediction table to the twitter_archive_clean data frame to get rid of the duplicated images at jpg_url column

16. Change the data type of tweet_id from int to string

Data Storage

I save my clean data in csv and named it twitter_archive_master.csv

Conclusion

In this project we completed the following:

1. Gathering the necessary data to explore the WeRateDogs twitter account including:
    - a file of the tweet archive data with tweet text and dog ratings
    - a file of image predictions based upon the photos tweeted
    - retweet and favorites counts gathered using the provided material
2. Assessed the quality of the above data, identifying quality and tidiness issues
3. Cleaned the above data based on the issues identified during the assessment
4. Analyzed the data to identify interesting insights, creating visualizations as necessary.

Source

Stack overflow

wrangle_act (maleina.com)

https://github.com/ahmed-gharib89/wrangle-and-analyze_data/blob/master/wrangle_act.ipynb