Wrangle and Analyze Data Report

WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Name: Mohammed Khalid Aldamadi.

Introduction:

We need 3 important steps for wrangle and analysis which are:

Gathering

Assessing

Cleaning

And now we will talk how we will use these three steps.

Gathering:

First, we will open a csv file (twitter_archive_enhanced.csv) using panda which talk about The WeRateDogs Twitter archive.

Second, we will open a tsv file (image_predictions.tsv) using panda which talk about The tweet image predictions.

And taking help by following this URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Finally, each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assessing:

We will use some functions on python to help us read the data and see what is wrong to fix it and we will divide it to two important types which are quality and tidiness.

Quality

tweet_id in tw table sholud be object type not integer.

Make the size of tweet_id 18 digits.

Favorite and Retweet in tw1 table sholud be integer type not float.

Timestamp not in good format.

Delete duplicated values in tw table.

Drop columns with missing values (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp).

Replace the incorrect names with None.

Fix the problem in rating numerator.

Multiple dog stage.

Tidiness

Add new colmun for count of (doggo, floofer, pupper and puppo).

Drop doggo, floofer, pupper and puppo columns.

Merge the three data sets in one.

Cleaning:

We will show how we will fix the problems above in cleaning section after we copies our tables to other tables:

Quality:

1) Define

Convert the tweet_id column data type in tw table from a int to a string using astype.

2) Define

Make the size of tweet_id column 18 using pad.

3) Define

Convert Favorite and Retweet columns data type in tw1 table from a int to a float using astype.

4) Define

Change the format of timestamp in tw table using strftime.

source: https://stackoverflow.com/questions/56698521/can-only-use-dt-accessor-with-datetimelike-values/56698574.

5) Define

Delete duplicated values in tw table by drop_duplicates method.

6) Define

We will drop these columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) in tw.

7) Define

replace all incomplete or incorrect name in the dog name column with (No Name) in tw table.

8) Define

Fixing the problem by turn all values that are above 10 and make it using loop and if condition in tw table.

9) Define

Handle the None value and merge columns for multiple stage.

Tidiness:

1) Define

Add new colmun for count in tw table using astype size.

source: https://stackoverflow.com/questions/17995024/how-to-assign-a-name-to-the-a-size-column.

2) Define

We will drop doggo, floofer, pupper and puppo columns from both tables then merage tw_clean1 table to tw_clean.

source: https://stackoverflow.com/questions/44327999/python-pandas-merge-multiple-dataframes/44338256.

3) Define

Merge the three data sets (tw, im, tw1) in one by outer join.