wrangle_report

February 10, 2021

0.0.1 WeRateDogs - Twitter Data

1. Gather Data I looked at the instructions given by the Udacity team on how to gather data for this data warngling analysis.

- I downloaded the data twitter-archive-enhanced.csv from Udacity.
- Next I downloaded the file image predictions file which is in the tsv format.
- Then I tried creating my twitter developer account but didn't get approval from tweeter so I downloaded the tweeter_api provided by udacity.

Once I had all the above three files, I created them into 3 different dataframes which are shown below.

- archive - this is a dataset "twitter-archive-enhanced.csv" which was converted into a dataframe and gives information on basic tweet data.
- image_predictions - This dataset will contain information about predictions about the image
- tweets_api - This dataset will contain information like tweet_id, no of retweets and no of favorites etc.,

0.0.2 2 Assessing

- In the assessing step, I tried to gather some quality and tidiness issue from the data I gathered.I assessed all files concurrently for similar quality or tiiness issues

Below are the quality and tidiness issues I accessed #### QUALITY

Archive

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_user_id, expanded_url and retweeted_status_timesttamp have missing rows
- timestamp and retweeted_timestamp in object form
- Correct denominators other than 10.
- Source columns have HTML tags
- tweet_id is in int type in both archive and image_prediction

- Rating_numerator with decimal values incorrectly extracted
- incorrect dog names contained(a, an, just,infugurated....etc) ##### Image_prediction
- dog breeds inconsistent,contains underscores, and have different case formatting
- Rename Columns(p1,p2,p3,p1_dog,p2_dog,p3_dog,p1_conf,p2_conf,p3_conf)to improv clarity #### Tidiness ##### archive_df table
- the columns doggo, floofer,pupper and puppo should be variables in a column dog_stage ##### All tables
- All three tables share the column tweet_id and should be merged to archive_df

0.0.3 3 Cleaning

For cleaning all the 3 dataframes, Below are the steps I took to clean the dataframes from observtions made in the accessinng phase

- I made a copy of all dataframes
- I joined the 3 dataframes based on their tweet_id
- I converted the datatype of "tweet_id" into string
- I created a column called dog_stage for the (puppo, pupper, doggo and the floofer stages), replaced "none" with null and dropped the null rows
- I dropped all duplicates including the ones as a result of creating the dog_stage column
- I converted timestamp which was in the string format to the datetime format
- I removed the underscores from the dog breeds and removed the inconsistency in the name format which was a mixture of both lower and upper case
- I removed the "><" from the source column so that the information would be properly extracted
- I made all rows in the rating_denominator 10 to remove row information that contained >10 or <10 since 10 is the only rating denominator
- I converted the ratings_denominators that are in decimal to float datatype
- Rename Columns(p1,p2,p3,p1_dog,p2_dog,p3_dog,p1_conf,p2_conf,p3_conf)to improve clarity
- I changed incorrect dog names (a, an, just, atually,all....etc)to none and then to nan and dropped the null rows
- I Rename Columns(p1,p2,p3,p1_dog,p2_dog,p3_dog,p1_conf,p2_conf,p3_conf)to improve clarity
- I removed the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,
- retweeted_status_user_id, and retweeted_status_timestamp columns which had so much missing rows as they won't give accurate results if analysed
- I also removed the retweet columnn since it's associated with the retweet_status_id

0.0.4 Storing Data

- I stored the final dataframe into csv file with name twitter_archive_master.csv with final data of 2060 rows and 21 columns

```
In [ ]:
```