Wrangle Act Report

August 23,2018

Data Gathering:

Data is gathered from three sources 'twitter_archive_enhanced.csv',' image_predictions.tsv' and 'tweet_json.txt.

First data was collected from the "twitter-archive-enhanced.csv" file which was in the same directory in which project notebook was located. The csv file was imported into pandas dataframe. The dataframe was named "df"

Second data was extracted programmatically from a URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv.Python's request library was used to extract data from this URL. This file was imported as a dataframe in pandas using tab as the separator. The dataframe was named "img_pred".

The third data was extracted from Twitter API using python's tweepy library. I needed to extract id, the favourites and retweet counts for each tweet. This data was then saved as a JSON file using UTF-8 encoding.

Data Assessing:

Data was assessed visually as well as programmatically. Through pandas info(), head(),describe() methods I was able to detect some quality and tideness issues like many tweets were retweeted, dataframe contained the faulty names,there were several empty values in in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

In several columns, the null values are treated as non-null values. Some entries contain "Nan" as string.

Some of the ratings did not look right because the expected value for numerator and denominator was around 10, but there were many values above 100 also.

Dog "stage" variable were in four columns: “doggo, floofer, pupper, puppo” individually. That could be condensed in one column.

Data Cleaning:

The copy of all the data frame were created as “df_clean”, “img_pred_clean”, tweet_info_clean. Retweets were removed and tweets which did not include images were also removed because those tweets were not dog ratings.

dog_stage column was created which showed the type of dog(dog stages).
Duplicate jpg_url in “img_pred_clean” dataframe was dropped . The text in 'img_pred' was made consistent and pretty.

Incorrect dog names in “df_clean” were changed. tweet without rating was removed. The value for rating numerator and denominator was corrected. The column “source” was made in more readable categories.

Datatypes of timestamp,dog_stage and tweet_id, in_reply_to_status_id,in_reply_to_user_id was changed to datetime, categorical, and to strings respectively.

After doing all the process of cleaning all the dataframe was merged in single data frame named “df_merge” and was stored in a csv file named “df_master.csv” and then dataset was visualized in different forms