WeRateDogs Data Wrangling Project

Data Gathering

What you have done, The data resources used, their formats and the technique used to gather each.

Twitter Archive ,Image_predictions gathering and additional data via Twitter API:

- I downloaded "twitter-archive-enhanced.csv" file manually and read it into a dataframe using pandas library
- I downloaded 'image_predictions.tsv' file programmatically using the Requests library and the provided url
- For twitter API I used the shortcut provided files and read the tweet_json.text line by line then extract 'tweet_id', 'retweet_count', 'favourite_count' each as a python dict and finally all those dicts applied in a list.

Output:

- archive_df
- image_prediction_df
- api_df

Data Assessment:

Visual assessment:

Excel

programmatic assessment:

- .info()
- .describe
-

Output:

Assessment Summary:

Quality aspects:

- archive_df:
  1. Dropping retweets

  2. Dropping (in_reply_t0_user_id , in-reply_to_status_id,retweeted_status_user_id,retweeted_status_timestamp,retweeted_status_id) columns

  3. posts without images

  4. Reprecentation of "null" values as a string "none" in 'nme' column

  5. dogs classification replace none value with ''

  6. tweet_id as int not string

  7. timestamp as object not datetime

  8. dropping dogs classification four columns after compining

  9. replace '' with null value in 'dog_stage' column

image_predictions_df:

Removing retweets

Api_df:

Removing retweets

Tidiness Aspects:

archive_df:

merging the four columns of dog stages under one column named ‘dog_stage’

api_df:

this dataframe does not show complete information on its own and it is better to be joined with archive_df

Data Cleaning:

** First we made copy of each dataframe

Data Cleaning:¶

archive_df

1. Dropping retweets and posts without images

Define:

dropping retweets and tweets with out images using:

- (in_reply_t0_user_id , in-reply_to_status_id,retweeted_status_user_id,retweeted_status_timestamp,retweeted_status_id) columns to drop retweets by re moving all not null values in those columns
- image_prediction_df to drop tweets with out images

2. Dropping (in_reply_t0_user_id , in-reply_to_status_id,retweeted_status_user_id, retweeted_status_timestamp, retweeted_status_id) columns

Define:
dropping those columns as they are no longer needed

3. Reprecentation of "null" values as a string "none" in 'name' column

Define:

In 'name' column we will replace the missed data from 'none' string to 'nan' value

1st we will replace the 'none' string with''

then we replace the empty string'' with the NAN value

4. dogs classification replace none value with empty string ''

5. tweet_id as int not string

Define:

convert tweet_id to string

6. 'timestamp' as string not date time

Define:

convert 'timestamp' to datetime

****** Dogs classifcation in more than columns(Tidiness issue)***

define:

dogs are classified under four columns so we are going to combine them under one column named dog_stage

7. Dropping the previous four classification columns

DEFINE:

dropping the four classification column after combining them under dog_stage column

8.Replace '' with null value in 'dog_stage' column

image_prediction_df(Quality aspect)

Removing retweets and replies ids

Define

dropping retweets using archive_df_clean ‘tweet_id’ column

api_df(Quality aspect)

Removing retweets

Define:

dropping retweets using archive_df_clean ‘tweet_id’ column

Join both archive_df and api_df(Tidiness issue)

Define:

creating a new dataframe compine both the archive_df and api_df

Output:

- twitter_archive_master
- image_clean_df