wrangle_report

September 7, 2022

0.1 DATA WRANGLING REPORT (WeRateDogs tweet Data Archive)

0.1.1 Introduction

In this project, we gathered three related datasets on Dog rating according to WeRateDogs tweets. They include an archived WeRateDogs tweets in csv format, image prediction carried out on the archive dataset in tsv format and tweeter API json data querried with the tweets IDs using tweepy. we accessed, cleaned and merged the data we gathered in order to gain insights from them.

0.1.2 Data Gathering

The archived WeRateDog tweet csv data, image prediction tsv data and tweeter API json data querried with tweet_id feature of the archived data using tweepy were gathered. icked only tweet_id, retweet_count and favorite_count data from the data returned from Twitter API though we saved the full data gotten from the Api as tweet-json.txt. we assigned the archived tweets .csv data to twitter_archive_df, the image prediction .tsv to image_prediction_df and the Api data (tweet_id, retweet_count and favorite_count) which was initially saved as tweet-json.txt to api_df.

0.1.3 Accessing Data

The data was accessed with different pandas methods (.info(),.describe(),.sample(),.unique()) and descoverd qaulity and tideness issues in twitter_archive_df and image_prediction_df.

0.1.4 Quality issues

twitter_archive_df Missing values in in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp and expanded_urls features

Tweet_id 835246439529840640 will throw a zero division error, the rating_denominator is zero.

We don't have an actual rating feature or column.

Source feature has 4 unique values which are string representation of an HTML anchor element. Both the anchor link and text has little or no insight to offer.

timestamp feature is an object instead of DateTime data type

The rating_denominator has some outliers, we are going to ignore this.

745 dogs have no names(None) while names like (a,an,by,his,old,my,O,such,not,one,this,all,the) given to some dogs are likely to be errornous

in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id and retweeted_status_user_id featurs has a float datatype instead of int

image_prediction_df samples of non dog predictions like desktop computers #### Tidiness issues ##### twitter_archive_df Entries of both retweets and reply; (in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp) features

Multtiple columns: [“doggo”, “flooter”, “pupper”, “puppo”] for one “stage” column

image_prediction_df: Multiple prediction columns; p1,p2,p3

0.1.5 Data Cleaning

We created a copy of the three dataframes before cleaning

twitter_archive_df We first cleaned the tidiness issues by filtering out all rows and columns associated with retweets and reply then we created a stage column from doggo, flooter, pupper, puppo colunms and droped them.

After cleaning the tidiness issue we discovered the row the rating_denominator is zero was already cleaned and the number of missing values reduced to just three which we dropped and we no longer had to worry about the wrong datatype of in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id and retweeted_status_user_id feaatures. we went ahead to create a rating feature (rating_numerator/rating_denominator) and drop the rating_numerator and rating_denominator features. We then convert timestamp feature to datetime and later renamed it to date. finaally,we replaced (a,an,by,his,old,my,O,such,not,one,this,all,the) with None

image_prediction_df first we filtered out non dog predictions and we picked p3 which has the highest confidence score and droped p1 and p3 with their confidence score. We then rename p3 to predicated_specie and p3_config to prediction_confidence.

0.1.6 Merging and Storing

We merged the three cleaned dataframes and assigned it to master_df variable. We later saved the master_df to our working folder as twitter_archive_master.csv