wrangle_report

August 15, 2022

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

First and foremost I gathered my data by downloading the twitter_archive_enhanced.csv file directly, I downloaded the image_predictions.tsv file programatically and I downloaded the tweet_json.txt file because I was denied api access by twitter. I then read the tweets from the json file and saved them in a csv file called "tweets.csv". Next loaded my gathered data into the workspace where i began to assess them both visually and programmatically, I came up with different errors and problems including but not limited to wrong data types, unnecessary columns, wrong data entry and so on. The next step involved was cleaning my data. I was able to convert the timestamp from string to datetime datatype with the use of the pandas to_datetime method. The source column had the a tag and I had two methods of cleaning this up. I could either use beautifulsoup to convert the string into a soup object and the access its content with the href attribute but i went with the use of regex to extract the url from yhe html a tag. I then searched the data for any rating denominators of zero, which I replaced with the correct rating by again using the amazing power of regex. Next with a bit of observation I discovered that missing values were represented with the string value of "None". I rightly assumed that this would lead to unforseen complications in the future and so in order to avoid this I replaced them with the NaN value by utilizing the np.nan method. the expnade_url from the archive_data table was missing some url and so I derived the missing value by combinng the url from the source column and the tweet id from the tweet_id column. The retweeted id and tweet id columns were integer and since these columns are supposed to contain the identification of tweets and should not be subject to any mathematical computations I converted them into strings. I used regex to derive the names which were wrongly labelled as "a" in the data. for those dogs whose names were actually present in the text column, I used regex to extract and replace the 'a' with the correct names, but for those whose names were not present I simply left them as they were. There were retweets in the data and even though these retweets might have diffent scores, there are still retweets of the sae dog and so result in the duplication of rows in the data. Through visual observation of the data, I got to discover that retweet columns have their text cells starting with "RT" an abbreviation for retweet, I then employed the poer of regex to filter out these rows wich I removed. I then faced the tidiness issues. I noticed that the four dog stages columns could be combined into a single table. I proceede to carry this out by first replacing nan with empty string characters and then adding the columns together. This created another problem because on closer observation i noticed that my newly created "stages" column had some combined stages even though I believed this could not be possible. On diving in to see the cause, I saw that the problem arose when tweets were referencing two dogs in a picture. Using regex, I filtered these entries out, I separated them into

two entries and then put them back into the dataframe. Lastly I believe that every information present in these three tables should actually belong in a single table that can all be linked by their tweet id's and so I merged the tables into one.