wrangle_report

September 3, 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.

My wrangling effort consisted of three main phases which are gathering,assessing and cleaning data which under this three main phases i was able to gather,assess both visually and programatically to spot a few tidiness and quality issues which later on were dealt with in the cleaning phase.

In the gathering phase, the first dataset('twitter_archive_enhanced.csv') i downloaded using the requests library same as the second dataset('image_predictions.tsv') while the third i used Tweepy to query Twitter's API having registering for twitters developers account which i was granted access to do so saving the third dataset in Json format('tweet_json.txt') which i had to read in line by line into a dataframe.

The second phase had me both visually using jupyter labs and programatically assess the data which using both methods i was able to put out at least 10 issues with the data sets both quality and tidiness isuess which i would list below:

Quality issues 1.the denominator should only be equal to 10

2.Not all tweets are dog tweets and some are retweets

3.Not all tweets are dog tweets some are replies

4.timestamp is in string format

5.id column name is different in api_tweets dataframe

6.dropping "img_num" column from 'image_pred' dataframe as it has no purpose

7.dropping deleted tweets from twitter_archive dataframe using ids from deleted_ids

8.dropping invalid names in 'name' column and replacing 'None' with Nan values

Tidiness issues 1.doggo,floofer,pupper,puppo should be under one column name 'dog_stage'

2.the three dataframes should be merged into one dataframe for better analysis

The third phase has me cleaning up all the above issues which i listed out from renaming columns to changing data types to dropping columns and rows to finally merging all the three datasets into one and saving it ''twitter_archive_master.csv",which further visualisations were carried on it.