wrangle_report

September 5, 2022

0.1 We Rate Dogs-Data Wrangling Project-Wrangle Report

In this report i will be describing my wrangling effort used to generate the master dataset for this project. The data wrangling process is divided into 3 sections known as - Data gathering - Accessing - Cleaning.

0.1.1 Data Gathering

The dataset for this project was gathered from 3 different source all in various formats.The steps i took for the data gathering phase is listed below:

> I started by importing all the necessary library needed for project execution. These includes * Pandas * Numpy * Request * Tweepy * Os & re * Json * Matplotlib & Seaborn

and thereafter proceeded to gather the datasets.

First Source This dataset 'twitter_enhanced_archive'was the file on hand, which was downloaded and uploaded on the jupyter notebook. Using Pandas read_csv function, i loaded this dataset into a dataframe called WRD_df.

Second Source The dataset was downloaded programmatically using the reequest library from udacity servers. I created a folder to store this file,then using requests to request for the file using URL,got a response of 200 indicating it was sucessful. Using pandas read function, i loaded the tsv file into a dataframe called imagepred_df.

Third Source I needed more information such as retweet count and favorite Count for the tweet ids in the twitter enhanced archive and this was to be gotten from twitter API. I started off by appling for a developers account which was approved and i generated the necessary keys for access. With this i did a sample tweet extract using one of the tweet ids and it was sucessful,then i proceeded to create a for loop to extract the parameters needed for this respective tweet ids,most were sucessful but i encountered 29 tweet ids which failed. The total duration of the extract was approximately 32mins. I created a dataframe(tweet_df) from the resulting dictionary.

0.1.2 Data Assessment

The 3 datasets were assessed visually and programmatically for Quality and Tidiness issues. I employed the use of excel for visual assessment and also manual scrolling on jupyter notebook, while i employed the use of functions such as head,tail,info,describe,regular expressions,duplicate,nunique etc for the programmatic assessment. Below is the summary of the Quality and Tidiness issue observed

Quality issues Summary Twitter Archive

1. The table contains 181 retweet as shown in retweeted_status_id.
2. Null Values recorded as None as seen in name column
3. Erroneous datatypes for tweet_id,timestamp,rating_numerator and denominator.
4. The dog stage name is floof instead of floofer shown in columns
5. The rating numerator and denominator contains zero values and certain incorrect values when they are decimals
6. Some columns containing null values such as retweeted_status_id,retweeted_status_user_id,retweeted_stat which needs to be dropped if not needed for analysis
7. The source contains an Html formatted string instead of the name of the utility used to post the tweet
8. The name column contains incorrect information such as a,None,an,the etc and all starting with lower cases
9. Some rows in expanded-url column has more than one url

Image Predictions

10. The tweet id is an integer instead of a string.
11. The image prediction table contains only 2075 rows hence others are considerd missing.
12. The image predictions columns names are not descriptive.
13. There are some strange values in the p1,p2 and p3 columns such as “orange, boxer,cardigan, car_wheel, can_opener”e.t.c , these cells must be checked first then deleted if not needed.
14. Inconsistency in the dog breed names, some starting with lowercase and some with upper case

Tidiness issues

1. Dog stages is represented in four columns.It is better to create one column for dog stages that contains the values (doggo, floofer, pupper and puppo).
2. The prediction,confidence and dog represented in 3 columns each
3. The three dataset all has information related to rating dogs hence merging is required.

0.1.3 Data Cleaning

This stage involves addressing the issues listed above using the Define,Code and Test framework. As best practice i created a copy of each dataframe before addressing the issues above. Firstly i removed the retweets as original tweets only was required as stated in the Project Motivation section,using numpy replaced none with nan,replaced an incorrect dog stage name,dropped columns not necessary for my analysis,extracted the rating numerator and denominator using

regex, merged the prediction dog breed column based on the prediction with the highest confidence(p1),merged the individual dog stages into one column,replaced incorrect dog name entries with nan,replaced html structure of tweet sources with the text components then proceeded to merge this datasets into one. On the merged dataset, i performed further cleaning such as extracting the correct urls from the text column and changing columns with erroneous datatypes.

0.1.4 Storage

I stored the cleaned merged dataset to a csv file called twitter_archive_master.csv.