wrangle_report

June 29, 2022

0.1 Reporting: wragle_report

Wrangle report There were three datasets that I used to finally create the final master clean dataset (twitter_archive_master.csv). this is a combination of the three datasets used in this exercise. The datasets include, twitter_archive_enhanced.csv, tweet-json.txt and image-predictions.tsv. Firstly, I made copies of all datasets in the project to use in the cleaning exercise. The issue identified and cleaned or handled include: 1. Changing the numerators from 0 on the twitter_archive_enahcend dataset. I did this by adding "10" to all values that were less than 10 since the rating method on weratedogs is the numerator being above 10. 2. Changing the datatypes of tweet_id on twitter archive dataset. I used the astype() method to achieve this. 3. I changed the values identified as None to Null which is logical and conventional naming of null values on dataFrames. 4. some names of dogs were unidentified sine they used pronouns, prepositions and some used articles e.g., the. While some could be misspelling of the names sch as "the" while the name could have been "Theo", I decided to change all these to no_name for simplicity. 5. I changed the datatype of time variables to datetime. I did this using the astype(datetime64) method on the specific columns of the twitter archive dataset. 6. In the image prediction dataset, the case used for the different 'p's values were inconsistent and therefore I used the str.title() method to change them to same case with the first letter of each capitalized. 7. I dropped unnecessary column such as retweet id status, source and reply related columns. This is in an effort tot only retain the columns that can be used in understanding the dataset better or generating insights. 8. I renamed the 'conf' in the image-prediction dataset to help with understating issues. The conf could be misinterpreted so I made it confidence using the rename function on the dataset. 9. The dog stages were split into different column which made the data untidy. I created a new column, extracted the dog stage from the text and assigned it to the new column using str.extract. on the twitter_archieve dataset 10. I reamed the id_str on the tweet_count dataset to twet_id using rename method on the dataset. 11. I changed the tweet_id datatype on the image prediction dataset to string type. This was to help with the next step. 12. I merged the three datasets to form one master dataset with the cleaned data and unnecessary columns dropped. This new merged master dataset was stored as twitter_archvie_master.csv. PS. I had an issue with the developer twitter account and opted to use the provided tweet-json.txt dataset. I will however follow up on the use of tweepy and use it later somewhere.