wrangle_report

June 23, 2022

0.1 Reporting: wrangle_report

0.1.1 Data wrangling involves gathering data from different sources, assessing them for quality and tidiness issues and then cleaning the data to ensure it is reliable, consistent and credible.

0.1.2 Gathering:

-This project involved gathering three datasets from three different sources viz: Twitter-archive.csv from the workspace using pd.read_csv Image_prediction table from webpage using response.get Tweet-json file from twitter API using tweepy.

0.1.3 Assessing:

Afterwards the three datasets were assessed visually and programmatically to spot inconsistencies in content and structure what we refer to as quality and tidiness issues. And the following issues were found:

Quality issues

Archive

1. Massive NaN values in the following columns; reply_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded urls.
2. Missing dog names has values of 'None', 'a', 'an', 'o'
3. Erroneous data types (all _id columns should be strings and timestamp should be datetime)
4. Name column not consistent with other dataframes.
5. Source values not presented properly
6. Rating denominator less than 10

Image_prediction

1. Erroneous data types (tweet_id should be a string)
2. Hidden dog breeds
3. Missing id's there are '2075' instead of '2353'

tweet_count

- No issues found.

Tidiness issues

1. Categorical data in different columns (doggo, floffer etc should be in one column)
2. Contains irrelevant columns
3. Column name id_str does not match with other tables
4. Irrelevant columns in image prediction table

0.1.4 Cleaning:

After spotting all the issues above, the cleaning process began witheach issue addressed and tested separately. I replaced the stopwords with np.nan using .replace() Converted the column type using to_datetime and astype Made sure names in tweet_archieve_clean and image_pred dataset are on lowercases to ensure consistency. In the archive_clean table, I changed the html ampersand code from "&amp ;" to "&" in the text column I changed values from any number less than 10 to 10 I changed the name column from "id" to "tweet_id" Used a function to extract dog breed from image_pred I extracted the dog stages from the four columns and put it in new column 'dog stage' using 'extract' then dropped the real columns I removed unwanted columns using drop method I renamed the id_str column in the tweet_count_clean dataset to be in line with other id values using the rename function.

0.1.5 Storing:

After the cleaning process I then merged all three datasets together as twitter_achive_master and sved it as a csv file.

0.1.6 Analysis :

To analyze the data, I asked the following questions; 1. Which dog stage is the most popular among dog lovers? 2. Which dog names are more popular with dog lovers? 3. What is the relationship between favorite count and the dog rating? 4. Which dog has the most favorite and retweet count in the dataset? 5. What are the top five dog breeds on this dataset?

0.1.7 Visulizations:

I plotted a few charts to explain relationships between variables viz; Bar chart Horizontal bar chart Scatter plot Histogram

0.1.8 Insights:

1. Puppers are really popular among dog lovers as a large majority of the dogs are of this stage.
2. Charlie is the most popular dog name followed closely by Lucy, Cooper and Oliver.
3. Time stamp is skewed to the right as we have more tweets from 2015 and 2016, retweet count is also skewed to the right with more retweets ranging from 1- 5000, and just a handful from 5000- 20000 and a few around 40000-70000
4. This is also the case for favorite count. Rating denominator is different because it consists of the same value (10).
5. From the scatterplot, I could see that there is no correlation between the favorite count and dog rating as they are both move in different directions.

6. The top dog breeds here as seen above are the golden retriever, labrador retriever, pembroke and the chihuahua. These breeds are popular among dog lovers.