Untitled

September 17, 2020

0.1 Wrangle Report

- The dataset wrangled in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

0.1.1 The WeRateDogs Twitter project goals included:

- Wrangling the twitter data through the following processes:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on the data wrangling efforts and data analyses and visualizations

0.1.2 Gathering Data

- Gathering Data for this Project involved obtaining three different datasets from three different sources. Each one testing a different way of obtaining a dataset.
- The first was to download a file manually and be able to open a csv file. In this case the file was called twitter_archive_enhanced.csv and was the file consisting of the largest amount of data.
- The second was to be able to download a file programmatically using Python Requests library. The file contained image predictions on the breed of the dog coming from a neural network on some of the tweets already downloaded in the archive file. The file was in tsv format and tested your ability to open this type of file successfully.
- The final dataset tested your ability to query Twitter's API and use a Python library called Tweepy to obtain further data on the tweets in the archive file using the tweet id. The Tweepy library returned the data in json format, from which it was possible to iterate through and append data to a file as a list of dictionaries and then a pandas data frame. A copy was saved in csv format of the data frame created.

0.1.3 Assessing Data

- Once the data was gathered, I began to assess the data on both quality and tidiness issues.Each of the data frames were evaluaated visually and programmatically. The following quality and tidiness issues were observedthe retweet and during the assessment. #### Tidiness Issues

- Merging the three DataFrames into one master DataFrame.
- The columns (doggo, floofer, pupper and puppo) do not need to be separated. Each dog will be classified as one of these classifications. It is better to create one column for dog classification that contains the values (doggo, floofer, pupper and puppo). #### Quality Issues
- wrong data types (tweet_id need to convert it to string)
- wrong data types (timestamp need to convert it to datetime)
- Reducing the 3 columns that predict the Breed of the dog in the table then drop('img_num','p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog') columns as we no loger need them.
- remove the tweets that has no image.
- removed retweets since they are essentially duplicates of the actual tweets then delete the columns related to retweets from the master dataframefirst_archive_master_df.
- removed replies since they are not the actual tweets then delete the columns related to replies from the master dataframefirst_archive_master_df.
- in name column, there are several values that are not dog names, like 'a', 'the', 'such', etc. all of these observations have lowercase characters.
- Ratings with decimal values incorrectly extracted for example 9.75/10 Currently, the value 75 would be captured as the rating numerator.
- source data column is not clear.

0.1.4 Cleaning Data

- After the assessment, I cleaned the data through the following means:
  - Define, Code and Test
- Merging the three DataFrames into one master DataFrame.
- Make one column for dog stage (doggo, floofer, pupper, and puppo) by saving value ('None' if no dog stage given). Also record if there are multiple dog stages, separating by a comma.
- convert tweet_id to string
- convert timestamp to datetime
- Reducing the 3 columns that predict the Breed of the dog in the image.
- removing tweets that has no image.
- removing rows that have non-empty retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
- deleting the columns related to retweets from the master dataframefirst_archive_master_df.
- deleting the columns related to replies from the master dataframefirst_archive_master_df.
- deleting the columns related to replies from the master dataframefirst_archive_master_df.
- Change dog names in names column.
- Getting the decimal ratings by matching the text pattern "#.#/#" and save as new rating (only denominators had decimals)

- Shorten data in source column by using regex library

```
In [ ]:
```