wrangle_report

October 2, 2019

The data was wrangled by first reading in data from a CSV file containing twitter archive data. The provided Tweet_IDs were then used to look up tweets using the Twitter API. Keys and tokens were generated using a Twitter Developer account. Using these keys, the Tweepy library was used to gather data from @dog_rates. Specifically, the retweet and favorite counts were collected for each tweet in the archive. Finally, we wanted to import the TSV file containing the image predictions corresponding to each archived tweet.

Once all the data was gathered, data clean up was started. Data cleanup was started on the tweet_archive table. Several tweets had to be eliminated so that only original tweets were included. Therefore, both replied and retweeted tweets had to be removed. Dog names were then cleaned because the names in the file were not always congruent with the name in the tweet. The unique classes of dog_rates then had to be fixed for redundancy. Data formatting was then conducted so that values were of the correct type. Once values were of the correct type, expanded_urls were fixed so that all of the tweets had a source link. The img_pred table was then cleaned so that dog types all had the same letter case and that they had the correct separator for organization purposes. Finally, data formatting was conducted so that values could be utilized for analysis.

After the data quality was cleaned up, it was tidied so that it could be exportable. Thus, several columns were dropped. Once they were cleaned up, they were merged to a final master table and then exported to a CSV.

[ ]: