wrangle_report

August 19, 2022

0.1 Reporting: wrangle_report

The dataset I wrangled, analyzed, and visualized is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. Firstly, I manually downloaded the twitter archive dataset available. Next, using the requests library, I programmatically downloaded the image predictions tsv file from the url provided. I got the third dataset from twitter API through my twitter developer account. I got the API data into a dataframe, and had 3 datasets ready to be cleaned.

I started the wrangling exercise by fixing the quality issues in the Twitter archive dataframe. The 'expanded_urls' had some missing values; I dropped the columns with missing values. Some tweets were not original; they were either retweets or replies to other tweets. I found them through the reply and retweet columns and dropped them. Some names of dogs appeared incorrect in the archive e.g 'none', 'a', 'an', etc. Using regex, I was able to get correct names from the archive text. I also dropped the rest of the names that were still showing as 'none'. To have a meaningful dataset for visualization, I also changed the data type of tweet IDs in the three dataframes from float to object (string). Normally, the numerator rating, in this case, can be larger than the denominator. The denominator is 10 for all cases, however, I found some ratings not following this pattern, so I created a match string format for the right rating. This way, I fixed the incorrect numerator and denominator ratings. I changed the timestamp column data type from object to datetime.

Moving to the image prediction dataset, I dropped the duplicated 'jpg URL' entries, and fixed names starting with lowercase letters. For the Twitter API data, I changed the 'Create date' data type from object to datetime.

To make the data tidy, I collapsed columns of different dog types (doggo, floofer, pupper and puppo) into one column. Finally, I merged the three dataframes into one dataframe, got insights, and visualized some part of the data.