Wrangle Report

May 11, 2021

1 Step 1: Gather Data

The Data is gathered from 3 resources:

1. The We Rate Dogs Twitter archive. This was downloaded manually using Udacity servers. This data was put into a data set called twitter_archive.
2. The image predictions file. This was downloaded programatically via Udacity servers. This data was put into a data set called image_predictions.
3. Each tweets retweet count and favorite count, by using the tweet IDs, which gave me the chance to query Twitter API, for each tweet's JSON data, in a file called tweet_json.txt file, using Python's tweepy library. This data was put into a dataframe called tweet_df.

Once when all the data was gathered it as imported into a Jupyter Notebook.

2 Step 2: Assess The Data

The objective was to find eight quality issues, and two tidiness issues. In order to do this, all three data sets needed to be visually assessed an reviewed.

2.1 Step 2A: Find Quality Issues

Quality issues include completeness, validity, accuracy, and consistentcy. Some of the quality issues found here are: 1. The first issue that needed to be fixed is that retweets need to be dropped from the data set. 2. In_reply_status_id, in_reply_to_user_id, retweeted_status_id, and retweeted_status_user_id are incomplete. 3. Favorite_count, timestamp, and retweet_count are all incorrect datatypes. 4. Sources column can be removed 5. Some of the names feature 'a', 'an' and 'the' 6. Some of the names aren't capitalized. 7. The rating_denominator has values other 10 8. The rating_numerator has some outliers that should be dropped.

2.2 Step 2B: Find Tidiness Issues

Tidiness issues are ways to make the data set more compact and easier to interpet. This includes:

```
In [ ]: 1. Merging the datasets into one Dataframe.
        2. Combining the doggo, floofer, pupper, and puppo into one column
        3. Combining the rating_numerator and rating_denominator into one column
```

3 Step 3: Clean the Data

The first The first step of cleaning was to merge all three data sets into one frame, as this made it much easier to clean and analyze the data. It is called DF_Master. From this point we can go through each issue, and make sure to define the problem, fix it with coding, and then lastly test it to make sure it ran smoothly.

In [ ]: