We Rate Dogs Data

Data Wrangling – A Project

Nora M

Phase 1: Gathering Data

To get the project started, the "twitter-archive-enhanced.csv" and "image-predictions.tsv" files were downloaded, then an "image predictions" folder was created. 'Twitter data' got created using the tweepy package by granting access to and downloading JSON's Twitter data. To get the JSON data, a list of tweet IDs from the "twitter archiveenhanced.csv" file was collected and then saved in a "tweet-json.txt" text file.

After the query's execution and data recording the text file was read line by line and extracted for each tweet by using JSON, and added the data to an empty list where a pandas Data Frame from the dictionary list saved it as "twitter data."

Phase 2: Assessing and Cleaning Data

Below is a list that mentions steps to be done to solve each issue regarding the table's cleanliness and quality

1. Keep original ratings so no retweets that have images

2. drop unneeded columns

3. Correct numerators with decimals

4. Wrong dog names.

5. Some rows have more than one dog stage

6. Remove the now unused columns.

7. Source column is HTML, not a normal string

8. Some texts have a hyper

9. The twitter JSON table should be added to the archive table.

10. Doggo, floofer, pupper, and puppo are all stages for dogs and should be in one column

Phase 2: Storing Cleaned Data

The data has been organized and is ready for analysis.