WeRateDogs Twitter Archive - Wrangle Report

Faisal Khan

22 June, 2022

Data Gathering

Data was gathered from 3 different source and then merged and stored in a separate files. The sources description is as follows:

1. WeRateDogs Twitter Enhanced archive, manually downloaded from Udacity's website
2. The image predictions file, programmatically downloaded from audacity using python's request library.
3. Extra information about tweets from twitter via its REST API and using Tweepy library.

Data was loaded and store in these variables: twitter_enhanced_archive, image_predictions and downloaded_tweets.

Assessment & Cleaning

Dataset was assessed both visually and programmatically. During the process some following issues were found and addressed

1. All ID values were changed to string
2. Unnecessary columns were dropped.
3. Invalid denominator values were fixed.
4. None values were replaced with NaN.
5. Wrong or invalid dog names were found and fixed.
6. Consistency was brought to dog names.
7. Column names were renamed to descriptive names.
8. Multi valued fields were found and fixed.
9. Some other required fixings

After cleaning the data, the datasets were merged and stored as a csv (twitter_archive_master.csv) file.