wrangle_report

September 20, 2022

##
Wrangle Report

0.0.1 Introduction

This project demonstrates the data wrangling process for the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. In this report I will provide a brief description of the data wrangling techniques that were used to gather, assess and clean the dog twitter archive.

0.0.2 Gather data

The following files were gathered for the analysis:

The WeRateDogs Twitter archive - This file (archive.csv) was downloaded manually and consists of basic tweet data for 2300+ tweets from WeRateDogs.

The tweet image predictions - i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) was downloaded programmatically from Udacity.

Each tweet's retweet count and favorite ("like") count -This file (tweet_json) contains JSON data for each tweet indicating the retweet and like counts.

0.0.3 Assess data

The three files obtained in the gathering phase were loaded into individual Pandas data frames for assessment. Each of the data frames were evaluated visually and programmatically. The following quality and tidiness issues were observed during the assessment.

Quality issues:

archive table - Some name column entries are not names - Some denominators are incorrect - Remove unused dog stage columns - Remove retweets

prediction table - Dropped entries that have p1_dog, p2_dog, & p3_dog values set to false. These are not dogs of any kind. - Dropped duplicate jpg_url entries - Dropped unused img_num column

tweet_json table - Tweet ID 886267009285017600 does not have a valid URL - Remove retweets

Tidiness issues: - Merged twitter archive, image predictions and tweet_json tables - Convert data type of tweet_id in tweet_json table from object to int64

0.0.4 Clean data

The quality and tidiness issues were cleaned using programmatic techniques such as:

- Dropping unnecessary columns from the tables
- Removing rows that consisted of retweets
- Removal of rows with duplicate information
- Deleted rows that did not have any dog predictions at all
- Combining all three data frames into a single data frame