Wrangle Report

Introduction

The purpose of this project is to practice wrangling data from multiple data sources. The data was assessed, tidied, cleaned and then combined into a final dataset for analysis. The dataset used is the Twitter archive of Twitter account WeRateDogs. WeRateDogs rates images of people's dogs with a humorous comment about the dog. An additional dataset was extracted using the Twitter API in order to obtain favorite counts and retweet counts for the tweets in the Twitter archive. The third dataset being included in the analysis is an image predictions dataset where all of the images from the archive tweets were processed though a neural network in an attempt to classify the dog breed represented in each image.

Gather Data

1. Twitter Archive file was provided in csv format
2. Programmatically Downloaded image-predictions in TSV using requests.
3. Used Tweepy to download tweets from Twitter API in JSON format

Wrangle Data

I. Assess

1) Each dataset was assessed using .info()
2) Data was viewed using .sample() and .head()
3) Individual columns of interest were assessed using .value_counts()

II. Tidy

1) Removed unneeded columns from JSON Tweets and Twitter Archive.
2) There are multiple dog type columns in the archive. They were merged into one column dog_class.
3) Extracted source in archive from html.
4) Added a numeric rating_num column in order to compare ratings.
5) Added a true_dog column to store the best predicted dog breed name in image predictions.
6) Merged all 3 different data sets into one cleaned data set twitter_archive_master.

III. Clean

1) Converted ID in JSON dataset to object.
2) Extracted character count from display_text_range in JSON dataset.
3) Identified all invalid names(starting with lowercase or "None") and converted to nan in archive dataset.
4) Removed reply and retweeted data from archive.
5) Converted tweet_id to object in archive.
6) Converted timestamp to datetime in archive.
7) Fixed bad values in rating_numerator in archive.
8) Fixed bad values in rating_denominator in archive.
9) Set tweet_id to object in image predictions.
10) Set breed to all lowercase in image predictions.

Store Cleaned Data

Final cleaned data was saved in twitter_archive_master.csv, ready for Analysis and Visualization.

Conclusion

The source data had numerous issues that needed to be resolved before the data could be analyzed. Using Python and Pandas libraries provided an efficient method for pragmatically assessing and cleansing the required data.