WeRate Dogs

A Data Wrangling Report.

Introduction

This project is based on the collection and analysis of the Popular WeRateDogs Twitter account. It involved taking data from multiple sources and combining them to get a coherent picture of WeRate's history.

The project required three datasets, the first contained archived data from the Twitter account in CSV format named, twitter_archive.csv. The second data source was derived from the Twitter API. The final dataset contains the result of an image prediction model that was run on the images of dogs. The project can be broken down into the following three sections.

1. Gather Stage
2. Assess Stage
3. Clean Stage

Gather

1. **Twitter Archive**: The twitter archive was in a csv contained on disk. We simply use pandas' read csv method.

2. **Image Predictions**: This data was stored in a remote location as a Tab Separated Value file (.tsv). We downloaded this data with the request module and loaded this data with pandas read csv method, specifying that the data is tab separated.

3. **Twitter API Data**: We use the Tweepy library to have streamlined access to the twitter API. We have to create a twitter developer account and when the account is approved, we retrieve authentication details. After authenticating the data we use the `get_status` method to retrieve tweets from our archive. Specifying the tweet id.

Assess and Clean

Here, we found and addressed a number of data issues.

Problem	**Solution**
The dataset contains retweets and replies.	Filter out retweets and replies statuses that are not null.
The source column contains HTML format. We can exclude this.	We simply excluded this column.
The timestamp column should be a datetime column.	We convert timestamps to data with pandas `to date method.
Some numerators have values lower than 10. However, we know that WeRateDogs constantly use ratings higher than 10.	Here we observe that some of these errors arise from misextracting decimals. We correct this by writing a custom regex pattern to extract decimals.
Some dog names are missing or nonsensical	We ignore this as we don't use it for our analysis
Some of the images don't contain dogs.	We filter on the `p1_dog` column
Dog life stage (i.e Doggo, puppo, etc) should be in one column.	We replace "Bones" with empty strings. Then concatenate the lifestage columns into one.
Twitter data has an unnecessary column that contains an index.	We take out this column.
Some columns are not useful for analysis	We exclude such columns
Some columns don't have descriptive names.	We rename columns for analysis, so they are more descriptive.
Tweets are spread across three different datasets.	Merge these datasets