Wrangle Report

Introduction:

This wrangle report presents the data wrangling process conducted on the WeRateDogs Twitter dataset. The dataset consists of three main tables: Twitter Archive, Image Prediction, and Tweet JSON. The objective of this project is to gather, assess, and clean the data in order to analyze the popularity of dog breeds and common dog stages in the tweets.

Gathering Data:

The data was collected from various sources:

1. Twitter_Archive was found on the Udacity website.
2. Image_Predictions could be found on the Udacity servers but was obtained programmatically using the request library.
3. Tweet_Json was obtained programmatically using tweep. With tweepy, I was able to get the file using the twitter API,

Assessing Data:

Both visual and programmatic assessment techniques were employed to identify any quality and tidiness issues in the data. Quality issues refer to problems with data accuracy, completeness, and consistency, while tidiness issues relate to the structure and organization of the data. The identified issues were noted for further cleaning.

Quality Issues:

Below is an overview of the quality issues found.

Twitter Archive

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be converted to string.
- Timestamp should be converted datetime (remove +0000 from end of strings)
- Source column should be be simplified. Results should be as such: IPhone, Web, Vine & TweetDeck.
- Only need the original ratings without any pictures, retweets, or replies. Additionally, remove any related columns. The picture aspect will be addressed at a later stage.
- Abnormal rating_denominator values exceeding 10.

Image Prediction

- Only a single column with the most confident dog breed prediction

Tweet Json

- Many of the predictions are absurd results like banana, pot, assault_rifle, etc.
- There is 66 duplicated jpg_url results.

Tidiness Issues:

Below is an overview of the tidiness issues found.

All Tables

- Data should be combined from all three tables and combined into.

twitter archive

- A new column should be created to store all the different stages of dogs in one column.

Conclusion:

The data wrangling process for the WeRateDogs Twitter dataset involved gathering data from multiple sources, assessing the data for quality and tidiness issues, and cleaning the data to address those issues. The resulting cleaned dataset enables analysis of the most popular dog breeds, the distribution of dog stages, and other related insights. The cleaned dataset is now ready for further analysis and visualization.