Data Wrangling Report

Introduction

This document the efforts put into the data wrangling process for the WeRate Dogs project. In this project, I have wangled the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Gathering Data

I gathered data from 3 different sources.

1. **Enhanced Twitter Archive**: I downloaded this csv file from project details page. This dataset contained an archive of WeRateDogs tweets. Loading this data set was quite simple and straight forward.
2. **Additional information from twitter API:** This was the trickiest part of data gathering. I had to get the list of tweet ids from the twitter archive and use Python's TweePy library to get additional information such as retweets, favorite counts from twitter. I had to review TweePy documentation and research multiple rate options for optimal Twitter API performance. It took a long time to get back the results of the query and also some tweets were missing. It's most likely due to the original tweet or account being deleted.
3. **Image Predictions dataset:** I had to programmatically download the image predictions dataset from a hosted server. I was able to do this via an http request and stored it locally to a tsv file.

Assessing Data

The enhanced twitter archive that is the biggest of all three datasets, had multiple quality and tidiness issues. Image predictions dataset had few quality issues and the additional information obtained from twitter API did not contain much issues.

I assessed the datasets both visually and programmatically. I came up with the below list of initial issues that needed to be fixed.

Quality Issues

1. There are some tweets that are retweets/duplicates of the original tweets
2. There are few entries that are missing image urls.
3. Some dog names are incorrect like 'a'.'an' etc
4. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id are of data type float
5. timestamp, retweeted_status_timestamp are of datatype String.
6. Some ratings are incorrect such as 450/500,24/7 etc. Some of the ratings are for group of dogs and some are parsing errors.
7. Some values in dog prediction columns p1,p2,p3 have underscores. Eg border_collie
8. The column names p1,p2,p3,p1_conf,p2_conf and p3_conf are not intuitive

Tidiness Issues

1. Dog stages doggo, floofer, pupper and puppo can be combined to a single column.
2. Three separate datasets are not required. The datasets can be merged.

Cleaning Data

Some of the fixes such as data type conversion and renaming columns were simple and improved the quality of the dataset significantly. However, some of the issues that needed to be fixed were complex and I had to do several iterations to clean it up. Quality issue item 6 related to rating was a bit challenging.

In addition to fixing quality issues, I was able to reshape and merge data so the resulting dataset was tidy.

Conclusion

This project was helpful in mastering the data wrangling skills with not so perfect real world data. The dataset itself and the process of gathering the data was quite interesting. I was able to fix tons of real time data issues and was able to iterate Gather-Asses-Clean methodology.