Reporting: wragle_report

- A report briefly explaining my wrangling efforts.

Introduction

This report is based on the data from twitter handle 'WeRateDogs'. WeRateDogs takes various dog's images with keen interest on dogs. They frown at any image other than dogs. Their rating is special because they often have higher numerator rating than the rating denominator. There are four dog stages puppo,floofer,pupper,doggo.

The Wrangling process is divided into three parts, **Gathering, Assessing and Cleaning.**

Gathering: I used three datasets in this project. I downloaded the first data manually; I downloaded the second data programmatically from a URL; the third dataset is a JSON file loaded into my Jupyter notebook as a dictionary derived from Twitter API(Udacity)

Assessing: I assessed my data visually using Excel and programmatically from my notebook. During the assessment stage, I discovered nine (9) quality issues and two(2) Tidiness issues. Visual assessment made me see why many incorrect dog names were in the data. I noticed some fixed naming patterns from the text: ('named', 'name to', 'name is', 'this is'). From this pattern, I knew how to correct the incorrect names. I also discovered a trend in the text. Text containing 'We only rate dogs' were tweets that weren't accepted by the dog rating platform. I was able to discover the reason from the assessment stage, which is that some tweets contained images that weren't dogs, people posed with their dogs, and some dog breeds like 'Samoyed' looked so much like bears, and they were assumed 'Bears' in some cases. The issues discovered are stated below:

Quality issues

1. Missing values represented as None instead of Null in `twitter_achive`*table*
2. Column with high amount of Null and not important to our Analysis in `twitter_achive`*table*
3. `twitter_achive` Retweet rows are like duplicate, and won't be of importance to our Analysis
4. Erroneous datatype(timestamp) in `twitter_achive`*table*
5. rating_numerator and rating_denominator datatype should be float not int in `twitter_achive`*table*
6. duplicate tweet present indicative in (name, expanded_url and rating) columns in `twitter_achive`*table*

7. Logan rating_numerator is 75 but the correct rating is 9.75 in `twitter_achive` *table*

8. Invalid rating of 960 and 0(960/0) for rating_numerator and rating_denominator) in `twitter_achive` *table*

9. Incorrect names(e.g a, very) in name column in `twitter_achive` *table*

10. Columns not important to our Analysis in `additional_tweet` *table*

Tidiness issues

1. One variable(stages) in four columns (doggo, floofer, 'pupper, puppo) in `twitter_achive` *table*

2. Image_prediction and additional_tweet should be part of `twitter_achive` *table*

Cleaning: This was the last stage of the data wrangling. It was the easiest stage because the issues with the data were in the assessment stage. I handled incorrect data types, names, ratings, missing values, duplicate rows, and unnecessary columns in this stage. I also corrected the data structure using the melt function from pandas, and merged the three tables into a master data frame.