Project: Wrangle and Analyze Data - WeRateDogs Twitter Data

By: Almutairi, Muhanned

Introduction

This data-wrangling report is made of tweets from the WeRateDogs Twitter account. The data wrangling effort was done with respect to the quality and tidiness issue.

Data wrangling

After inspection of the data, several issues were indicated as below:

Quality

`twit_arch` table

- 1 - The timestamp field is in string format (object)
- 2 - Some of the rows have invalid strings in the name column, e.g. "a", "an", "in".
- 3 - Nan value in the name column.
- 4 - There are some tweets beyond August 1st, 2017.
- 5 - Erroneous data types in the source column are not consistent..html code
- 6 - Dataset contains retweets.
- 7 - Extra characters after '&'.

`image_pred` table

- 8 - image prediction shows that has three level of confidence and some image is not images of dogs.

Tidiness

- 1 - `image_pred` should be part of the `twit_arch` table.
- 2 - `twet_df` should be part of the `twit_arch` table as well.
- 3 - doggo, floofer, pupper, puppo columns in `twit_arch` should be combined into a single column as this is one variable that identifies the stage of the dog.

Cleaning

In the cleaning section, each issue was defined then the code was created to fix the issue then lastly the text code was performed to make sure the issue was indicated earlier was fixed. These steps were very helpful to engage readers. You may look at 'wrangle_act' Jupyter notebook for all code used.