Wrangle Report

Introduction

The goal of this project is to practice data wrangling, where there are different files are put together to give more insights. The data used here are from @WeRateDogs twitter account, all are three different files. Here we will gather these data, assess them, and then finally clean them up.

Gathering

Here we loaded three different files into the Jupyter notebook, two flat files and one Json file. The first file contains information like tweet IDs, time stamp, tweet text, dogs ratings, dogs names and their stage. The Image prediction file contains the tweet IDs, in addition to a prediction of the dog's breed. Finally a Josn file that contains additional information about the tweets, count of retweets and favorites.

Assessing

All gathered data had been assessed visually using Microsoft Excel, and programmatically on Jupyter Notebook for quality and tidiness issues.

Quality

- **twitter_archive file:**
    1. A number of retweets among the data.
    2. Nominator type should be float.
    3. Nominators like 9.5 and 13.5 are wrong (there are others).
    4. Denominators other than 10.
    5. Wrong dog names (mistaken for words).
    6. expanded_urls with no images.
    7. duplicated images in expanded_urls.
    8. timestamp in twitter_archive should be datetime.

- **image_prediction file:**
    9. Not all tweets in twitter_archive are in the image_prediction file (different number of records).

Tideness

- **twitter_archive file:**
    10. Unneeded columns (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) in twitter_archive.
    11. Dog stages has four columns.
    12. Gather all data into one dataframe

Cleaning

In the cleaning stage, all the points that were found are grouped based on the general issue. So we had nine main issues to be fixed, which are:

1. Removing retweeted data
2. Fixing wrong rating entries
3. Fixing dog names
4. Fixing images URLs
5. Fixing timestamp type
6. Merging dog stages to one column
7. Dropping unwanted columns
8. New prediction column with confidence level
9. New columns for number of retweets and favorites