Data Wrangling Report: WeRateDogs

By Patrick Bloomingdale

Date: December 11, 2018

This project is part of Udacity's Data Analyst Nanodegree

Introduction

For this project we were tasked with:

- Data wrangling, consisting of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Analyzing and visualizing the data that was wrangled
- Creating two reports:
  1. Data wrangling efforts
  2. Data analysis and visualization

Gather

We had to gather data from the following three sources:

1. twitter-archive-enhanced.csv: file provided by Udacity and downloaded manually
2. image-predictions.tsv: file hosted on Udacity's server and downloaded programmatically using Requests library
3. tweet_json.txt: queried the Twitter API for tweets in the Twitter archive using Tweepy library and saved JSON in a text file

Assess

After I gathered each of the different datasets, I assessed them visually and programmatically for quality and tidiness issues. I found the following issues:

Data Quality Issues

For the df_arch table:

1. name: has values of "None" instead of "NaN" and names that are not the name of the dog, such as:
   - “a", "actually", "an", "None", "not", "old", "Officially", "the", "this".
2. doggo, floofer, pupper, and puppo: have values of "None" instead of "NaN"
3. Some of the tweets in the dataset are retweet.
4. rating_numerator: numerators with decimals were not converted correctly. For example:
   - a numerator of 11.27 value was 27.
5. timestamp: column object not date
6. rating_numerator: column interger not float
7. created_at: column object not date

For the df_images table:

8. p1, p2, and p3 contain images that are not dogs

For the df_tweet table:

9. created_at: column object not date

For the df_arch table and df_images table:

10. df_arch_clean and df_images_clean tables contain columns that are not needed

Tidiness

1. Column headers (variables) are values, not variable names .
    - doggo, floofer, pupper, and puppo: These columns are the different dog stages (values) and should be stored in a single column.
2. Some of the same variables are named differently on the different DataFrames
    - tweet_id and id are the same variable
    - text and full_text are the same variable
3. The three tables (df_arch, df_images, df_tweet) should be combined into one table for analysis

Clean

I addressed the data quality and tidiness issues listed above and saved the combined dataframe as twitter_archive_master and stored the data in a csv file called twitter_archive_master.csv.