Wrangle Report

~ Wrangle and Analyze Data ~

Audrey S Tan

April, 2019

In this project, I applied the concepts learned from the lessons in Data Wrangling, gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then cleaned it. From the cleansed dataset, I went on to produce 3 data insights and visualizations. Below is the summary of the Gather, Assess, Clean, Analyze and Visualization steps I went through:

1) Gather data from three different sources:

- WeRateDogs Twitter archive. This is provided by Udacity in a csv file format and contains 5000+ basic tweet data about dog rating, name, and "stage".
- Tweet image predictions. This is also provided by Udacity in tsv file format which I downloaded programmatically from Udacity site. This file contains dog breed prediction results (from a Neural Network classifier) for every dog images from the WeRateDogs Twitter archive.
- Additional Twitter Data. The data resides on Twitter site and can be pulled via their API tweepy. I used the API to query additional data (in JSON format) and downloaded into a file named tweet_json.txt. This file has favorite and retweet count information for each tweet ID in the WeRateDogs Twitter archive, which are crucial for the dog rating analysis.

2) Assess data for quality and tidiness:

- I inspected the three datasets visually and programmatically to produce a list of quality and tidiness issues.
- Quality issues include:

  - various issues pertains to incorrect rating numerator and denominator values in the main twitter dataset.
  - inproper data types for tweet id, timestamp, rating numerator and denominator in the main twitter dataset.
  - invalid dog names and inconsistent dog naming convention in the main and secondary twitter datasets.
  - presence of retweet and reply-to data in the main and secondary twitter data datasets.
  - superfluous columns in the main twitter data dataset.

- Tidiness issues include:

  - dog stages span four different columns in the main twitter dataset which can and should be combined into one.
  - three types of observations (dog, non dog and partial) in the prediction dataset
  - the three datasets can be combined into one single dataset

3) Clean data to fix quality and tidiness issues identified:

- for each of the issues identified in each dataset, prescribed a code fix, built, executed and tested the code fix.
- combined the three datasets into a single master dataset, store it to a csv file and a python database table.

4) Analyze and visualize the wrangled data:

- looked at the cleaned master dataset and produce three insights with visualizations.
- Insights and visulizations produced include:

  - correlation between favorite and retweet counts.
  - the trend of favorite and retweet counts with respect to time and classification of dog species
  - performance of the dog image classifier

Created by Audrey S Tan.