Wrangle and Analyze Data Project

Wrangle Report

Introduction

In this report, I will describe my wrangling efforts on the WeRateDogs Twitter data. It includes three parts: gathering data, assessing data, cleaning data, and storing data.

Gathering Data

First, I downloaded “twitter-archive-enhanced.csv” manually from Udacity. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
Second, I downloaded “image-predictions.tsv” file using “Requests” library. This file contains the top three predictions of dog breed for each dog image from the WeRateDogs Enhanced Twitter Archive.
Third, I downloaded the Twitter API file manually from Udacity.
After downloading these data, I read them into a pandas data frame for the next step.

Assessing Data

I conducted both visual assessment and programmatic assessment. First, I observed the data in spreadsheet programs (i.e., excel, google sheets). I found the following issues.

- There is unnecessary HTTP in the source column.
- Dog 'stage' classification (doggo, floofer, pupper or puppo) is in multiple columns instead of one.
- Some of the gathered tweets are replies that should be removed.
- Some of the gathered tweets are retweets that should be removed.

Next, using various methods and functions in the pandas library, I was able to spot other quality and tidiness issues. I identified the following issues:

- Denominator of some ratings is not 10
- Numerator of some ratings is greater than 10
- Float ratings have been incorrectly read from the text of the tweet.
- We have 639 expanded URLs which contain more than one URL address.
- Source column has HTTP code that needs to be removed.
- 66 jpg_url duplicates were found.

Cleaning Data

Before I started, I created a copy of each data and performed cleaning on the copies. I then run code to fix each issue I assessed based on quality and tidiness rules. I always run a test to make sure the issues are fixed.

Storing Data

After cleaning data, I stored the data t a CSV file named “twitter_archive_master.csv”. I later used the data to perform visualization.