Saron Tessera

Data Wrangling Project

Wrangling, Analyzing and visualizing WeRateDogs dataset

Introduction

In this project, we will be wrangling, analyzing and visualizing the data set from twitter named WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. We will be using WeRateDogs twitter archives. These archives contain basic twitter information. Using the gathered data provided, we will assess, clean, store, analyze and visualize the data.

Data Gathering

We started by importing the necessary libraries like numpy, tweepy, json and matplotlib.

First, we make sure that we read the csv file that was provided by the Udacity course.

Second, the next part of the gathering was to download the image_predition.tsv file. This file contains the different URL, breed and so on.

Third, we must use the tweet ids in the we rate dogs to query twitter API for each tweet's json data using phyton's tweepy library and sore entire set to json data.

Data Assessing

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. we will detect and document at least eight (8) quality issues and two (2) tidiness issues.

The issues that I found are listed below

Quality Issues

- The Data contains retweets. It shows that there are 181 retweet that needs to be excluded**.**
- It's hard to read all the dataset, so remove unnecessary fields from the analysis
- Inconsistent naming of the dog's name
- timestamp and retweeted_status_timestamp are type 'object'
- p1, p2, and p3 contain underscores instead of spaces in the labels
- In the df data frame, there are 2356 but, in the images, it contains 2075.
- It contains 0 data

Tidiness issue

- Issues with the structure of the data
- Parse the datetime information into separate columns
- Drop columns that are not needed & rearrange column order for an easier read
- Combine each dog stage column into a single column named "stage"
- tweet_id column needs to be converted from a number to string value
- Date and Time columns need to be converted to datetime objects
- Rating columns need to be converted to float values

Cleaning Data

In order to clean the data, I first must remove all the retweets from the data frame. The next quality clean I did is to remove all the unnecessary column that are not used like ('retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'in_reply_to_status_id). I next remove the column like doggo, floofer, pupper, puppo to dogs_stage. We Change the datatype of 'timestamp' to datetime. I fixed the correct the ‘rating_numerator’ values from the text information.

Storing, Analyzing and Visualizing

We stored the information in a master csv file and after that we will analyze the stored information.

The final step will be to visualize the information.