Wrangle Report

Introduction

The dataset used in this project is the tweet archive of user @dog_rates, known as **WeRateDogs** that rates dog's with a humorous comment. In general ratings are from 1 to 10, but they thinks that every dog deserves at least 10 so their numerators are always greater than 10 with a denominator of 10. For eg,(12/10 , 14/10)

Project Details

- Data wrangling, which consists of:
  - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

Gathering Data

Data used in this project consists of 3 different datasets gathered from following sources:

1. The WeRateDogs Twitter archive @dog_rates. The twitter_archive_enhanced.csv file was provided by Udacity.
2. The tweet image predictions, i.e., what breed of dog are present in each tweet according to a neural network. This file was downloaded programmatically which was hosted on Udacity.
3. Twitter API and Python's Tweepy library was used to gather each tweet's retweet count and favorite ("like") count, friends count, source, retweeted status and url.

Assessing Data

- Jupyter notebook is used to convert all 3 datasets to data frames for analysis.

- Different functions were used like value_counts,info,describe,etc.

Cleaning Data

The strategy was used for cleaning the data is **Define, Code and Test.**

Following steps were used in the cleaning process:

- Merge the clean versions of twiter_archive, images_predict, and twitter_counts_df.
- Create one column for the various dog types: doggo, floofer, pupper, puppo
- Remove columns no longer needed: in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
- Delete retweets
- Remove columns no longer needed
- Change tweet_id from an integer to a string
- Change the timestamp to correct datetime format
- Correct naming issues
- Standardize dog ratings
- Create a dog_breed column.