WeRateDogs Wrangling Report

Introduction

The WeRateDogs wrangling project is a data wrangling project where I wrangled, analyzed and visualized the tweet archive of Twitter User @dog_rates (a Twitter account that rates people's dogs with humorous comments about the dog). This project consists of five steps:

1. Gather Data
2. Assess Data
3. Clean Data
4. Generate insights
5. Visualize outcomes

Gather Data

I gathered three files for the analysis. For this data gathering, I connected to Twitter's API to download data from Twitter. The three files I gathered are:

WeRateDogs Twitter archive Data: Renamed archive.csv, this data contains over 2000 tweets downloaded from WeRateDogs.

Image prediction data: This file was programmatically downloaded from Udacity. It contains image files of dogs or other images present with the tweets. It was renamed prediction.csv

Like and Retweet count data: Renamed tweet_json, this file contains like and retweet count for each tweet.

Assess Data

The three files obtained were visually and programmatically assessed to understand the data quality and tidiness issues they may have.

Quality issues

tweet_json

1. the date format in created_at contains different features that should be separated.
2. Tweet_json has retweets. Remove them

Prediction

3. Dog and conf have three different columns. Consolidate them into one
4. jpg_url column has 66 duplicate entries. Remove duplicates

Archive

5. Archive has numerous null values
6. The source column has a complex url. Remove the a href//https part

General

7. Prediction and archive columns have complex names.
8. tweet_id should not be an integer. Convert to string

Tidiness issues

1. The tables should be merged into tweet and images
2. Dog breeds/types have different colums. Consolidate them

Clean Data

The quality and tidiness issues were programmatically corrected. For each issue, I went through the **define, code** and **test** phase of data cleaning. These are (but not limited to):

- Removing small letters in the name column of archive.csv using Regular expressions
- Rewriting the source code link in the archive.csv file by replacing the untidy source code with clean html address
- Removing retweeted tweets in all the files to maintain tweet credibility and avoid duplicates
- Combining the three files into one file for tidiness and easy accessibility

Generate Insights

After cleaning the data, I further assessed the data programmatically to find patterns and generate insights. The three insights I noticed are:

1. Cooper is the most common dog name
2. Clumber has the highest mean numerator rating of 27.
3. Golden retriever is the most common breed

Visualize Data

For ease of communication, I visualized my insight using Matplotlib to communicate effectively what my results were.