Data Wrangling of WeRateDogs by Chinelo Okafor

Introduction

The dataset I wrangled for this project is WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about dogs. These ratings almost always have a denominator of 19. However, the rating is sometimes greater than 10, because, the dogs are good dogs. WeRateDogs has over 4 million followers and has received international coverage.

Steps to Data Wrangling

I followed the following steps to achieve wrangling of WeRateDogs twitter data. They are as follows;

Step 1 : Gather Data

- Downloaded the WeRateDogs Twitter archive file (twitter_archive_enhanced.csv) manually provided in the classroom.
- Programmatically downloaded the image_predictions.tsv hosted on Udacity's servers .
- Gathered through Twitter's API using the tweet IDs, retweet count and favorite count data and were stored in tweet_json.txt file.

Step 2: Assess Data

Using .info(), .describe() and .head() was able to assess the data to get the following quality and tidy issues

Twitter_archive

Quality issues

1. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, The variables above data is incomplete, all should be removed.
2. Filter out outliers in rating_numerator and rating_denominator

Tidiness Issues

1. dog stages (doggo, floofer, pupper, and puppo) should be values of a single variable.
2. Rename rating_denominator to "dog_rating"
3. Create unique columns from timestamp (Year, month, day, hour).

dog_predictions

Quality issues

1. Inconsistent dog names in p1, p2, p3 values
2. p1, p2, and p3 should be converted to title format.
3. Dog breeds are in both lower and Upper cases, should all start with upper cases.
4. p1, p2, and p3 should be replaced with space.

tweep_df

Quality issues

1. contributors, coordinates, extended_entities, geo, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, place, possibly_sensitive, possibly_sensitive_appealable, quoted_status, quoted_status_id, quoted_status_id_str, quoted_status_permalink, retweeted_status these variables have incomplete data.
2. contributors, coordinates, geo, place, no data available should be removed.

Tidiness issues

1. id should be renamed tweet_id.

Step 3 : Clean Data

The Quality and Tidy issues above was corrected and the three data merged for Data Visualization

Step 4: Visualize Data

Data is visualized in other to gather information on how the dogs are rated.