Project Wrangle and Analyze Data with WeRateDogs

Introduction

This project is about wrangling WeRateDogs Twitter data from three different sources. WeRateDogs is a Twitter account where dog pictures are rated and described with funny comments. The rates are almost always above maximum. I will go through the three data wrangling steps: Gathering, Assessing and Cleaning.

Gathering

The data for this project have been gathered from three different sources:

- file on hand: the WeRateDog Twitter archive (twitter-archive-master.csv), with information about the twitter-id, twitter-text, dog name, ratings and more.
- download programatically: a tweet image prediction file (image-predictions.tsv), hosted on udacity's server. This file is containing information about the recognized item on the tweet pictures.
- query of the Twitter API using Tweepy - which unfortunately didn't work for me, so the file tweet_json.txt has been used for this project. It contains a large range of tweets information. These three files have been uploaded into three different dataframes.

Assessing

The assessing was made successively for all three dataframes. Here are the main assessed issues:

- remove the rows which are retweets (Project instructions).
- change datatypes of several columns to str or timedata (quality issue)
- group some column information ("doggo", "floofer", "pupper" and "puppo") into one, because each variable should be in one column (tidiness issue)
- remove the unnecessary columns (quality issue)
- check the data entries of some columns for inconsistent data (dog name / denominator / numerator), correct the data of delete the entry/row (quality issue)
- create new columns with proportion calculation of ratings and most likely dog breed (tidiness issue)
- improve the data entries for dog names (capitalize the names) (quality issue)
- change column name (quality issue)
- merge the three cleaned dataframes into one (tidiness issue)

Cleaning

After creating a copy for all dataframes, I went through all the defined issues from the assessing part, cleaned these issues and tested the results. After making sure that the issues were solved, I merged the three dataframes to one (df_clean).
After this cleaning steps, I reassessed the column 'rating_numerator' and cleaned up some new defined issues.

Summary/Conclusion

After performing all the wrangling steps on all dataframes, I got a single one with cleaned data, which I used to analyze the tweet data and gain some insights.
I found that posts with a high number of likes and/or retweets are more likely to contain the following information: name of the dog, dog stage and a nice recognizable dog picture.