Wrangle and analyze data from WeRateDogs' twitter archive – internal summary report

In this project 2356 tweets from the @dog_rates account were wrangled and analyzed. This approach consisted of 4 distinct steps: gathering the data, assessing data quality and tidiness, cleaning and finally analyzing the data. All of these steps including final conclusions plus visualizations can be found in the Jupyter Notebook wrangle_act.ipynb.

Data were gathered from 4 different sources:
(1) the twitter_archive_enhanced.csv file was on hand at the beginning of the project;
(2) a .tsv file (called image_predictions.tsv) containing image prediction data for the same data set was downloaded from the internet using the get method from pythons requests package;
(3) the content of each of the tweets was retrieved in json-format through Twitter's API using the tweepy package and from these retweet_count and favorite_count were extracted;
(4) a list of all dog breeds was retrieved from Wikipedia by accessing the website and parsing the html-tags using the BeautifulSoup package.

The 4 data sets were loaded into 4 different pandas dataframes. During the assessment step 11 data quality issues were identified including incorrect data types for certain columns, incorrect usage of the NoneType value ('None' string used instead), and incorrectly extracted dog names and rating numerators. In addition, several data tidiness problems were detected such as the use of 4 different columns for the different dog stages instead of one.

Each of the data quality and data tidiness problems were subsequently addressed during the clean-up step. To make this process as transparent as possible to the reviewer every issue was first defined in words, then the code to fix the problem was run, and then a test was performed to ensure that the issue had been fixed. After multiple iterations a final data frame called 'master' was created and also stored as a .csv file and a sqlite database.

Finally, the cleaned data were analyzed and visualizations were created for three insights. The conclusions are summarized in the act_report.pdf. In a nutshell, we found that (1) dog ratings increased over time, (2) number of tweets posted per month decreased over time and stabilized at around 80 tweets per month more recently, and (3) the median number of favorite tweets steadily increased over time.