Wrangle and Analyze Dog Tweets Data

WRANGLE REPORT

In this project, I worked with three data sources: 1) the WeRateDogs Twitter archive, 2) the image predictions data, and 3) the additional tweets data. Data wrangling included gathering, assessing, and cleaning of all three datasets, and creating the master dataset by joining the three cleansed smaller datasets together.

Gathering

The WeRateDogs Twitter archive, available from the Udacity url in a csv format, was downloaded manually and loaded into a Pandas dataframe. The image predictions data, also available from the Udacity url, but in a tsv format, was downloaded programmatically using Python Requests library and loaded into a Pandas dataframe as well.

To obtain the additional tweets data, I have created a Twitter user and developer accounts, and a Twitter application, since the Twitter API requires authentication. Then, I used the Tweepy library to query the API - in each request, the tweet's id from the WeRateDogs archive was used. I have read each tweet's json into a file line by line. Afterwards, I used the Python json library to read from the created text file and created a Pandas dataframe with one row per tweet using columns tweet_id, created_at, retweet_count and favorite_count.

Assessing

The next step in the data wrangling process was assessment of data quality and tidiness for the three datasets. First, I explored the data visually to spot obvious issues such as column (variable) names and their contents, missing data and representation of missing values, unexpected (potentially erroneous) values in each column, data consistency, how the date and time is represented, among others.

Next, I assessed the data programmatically using common Pandas methods. I have identified both quality and tidiness issues to be cleansed in the next step (note: I focused on issues to fulfil project requirements, and to be able to perform further analysis and/or visualizations on the joined data).

Cleaning

In the last step of the data wrangling process, I have cleaned the identified issues one by one using the approach 'define - code - test'. I have used common Pandas methods to do replacements including regex expressions, column and/or row removals, merging of datasets, data types conversions and others.

As a first step, I cleaned the issues related to the completeness of data - I removed retweets, tweets without images, tweets without ratings and tweets not related to dogs at all. Then I continued with the tidiness issues - I prepared the Twitter archive and the image predictions dataframes to be merged into one master dataset. I decided to keep only information that I planned to analyze and visualize; and removed some data because of this. Although these additional data could provide valuable insights, for example the image predictions data could be analyzed to assess the used neural network model, I decided to constrain the dataset with respect to planned analyses and visualizations.

The master dataset contains only dog-related tweets with at least one image and successful prediction of the dog breed. The prepared master dataset was exported into csv.