Udacity Project:
Wrangling and Analyzing Data

by Stefan Cornelißen, 2018-01-23

Wrangle Report - WeRateDogs Analysis

Within the Udacity Project, a typical data wrangling process needed to be accomplished. This included a gathering data from a variety of sources, such as CSV-files, TSV-files and an API. The project aimed to analyze data from the WeRateDogs Twitter account. On this account pictures of funny or cute dogs are published and rated. In order to run an analysis to answer questions like "Which dog breed is achieved the highest rating, or which dog breed earned most favorites or retweets", three different sources of data needed to be combined.

Archive of the WeRateDogs Twitter Account

The main challenge here were missing values across several columns. The column expaneded_urls for instance, contained many missing values, although the image is very essential for the analysis. In addition, many URLs were present duplicately within one single cell. Next the numerator and denominator of the rating were each located in single columns. They were scraped that way from the Tweet text. Both needed to be combined to a single value. Besides that, this table had many tidiness issues. For example actual values were represented in column names. Lastly this table contained much unusable data from retweets and replies. These and some other issues were cleaned in process.

Image Predication Data

To make the analysis more valuable, the dog pictures of the Tweets were analyzed to predict the dog breed. This has been achieved using image recognition algorithm. Like the archive, this dataset was given. The main challenge here was to interpret the structure in order to be able to join the predictions with the ratings from the archive using the identifier "tweet ID". Furthermore the image prediction recognized things that were also present in the picture, like tennis balls, seat belts, … . Except from some duplicates and mixed capitalization the table had no major shortcomings.

Twitter Performance Data: Retrieving retweet and favorite counts using the Twitter API

In order to run the analysis on the "performance" of the dog pictures to the WeRateDogs audience, those metrics needed to be retrieved separately. To accomplish this task, I needed to run the concerning tweet ids through the Twitter API to retrieve the needed information. Within this process much more metrics and dimensions were delivered from the API. Hence, the wished metrics needed to be extracted from the JSON object to a separate table.

Final merging

In preparation for the analysis part, these three data sources/tables needed to be put in relationship with each other. The key element in each dataset is the tweet ID. During the wrangling process I ensured that these were unique, so that two connecting one to one relationships could be established. Since I decided to run the analysis using the Python Pandas library, the easiest way to establish the relationships, was to combine all three tables to one major table. To create a better overview, unneeded columns were dropped.