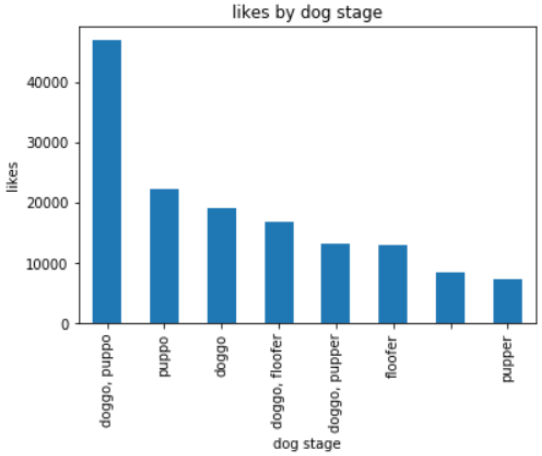
Analysis Report

The following analysis is based on the wrangling report where a dataset of nearly 2000 observations was gathered from three different resources: an exclusive dataset from the Twitter user WeRateDogs, a dataset that predicted the breed of dog in the tweet images with a neural network and basic tweet information gathered from the Twitters database. These datasets were merged and quality and tidiness issues were cleaned.

The first question that interested me, was which breed of dog might be the cutest and in consequence would get most likes. For this investigation a supposed that the first guess of the neural network was right, disregarding the confidence interval. According to the neural network “laptop” is the breed with the most likes (66,382), whereas “desktop_computer” is the breed with the feswest likes (104). There are 374 different "breeds" and apparently most of them aren't real dog-"breeds". In consequence, I filtered the observations by p1_dog == True to obtain only dog-pictures according to the first guess of the neural network. This reduced the observations to 111 "real dog breeds", with Saluki beeing the most liked breed on average. Gary (on the right), is the dog within the Saluki breed that got the most likes.

In the next step, I looked at which dog stage got the most likes.:



The most liked dog_stage is doggo, puppo, but since it only contains one observation and hence the mean of likes isn't really a fair indicator, I ignored it. The next best categorie is puppo containing 22 observations. And by far the dog stage with the fewest likes is pupper.

This result is really suprising, since I would have expected that younger dogs would be cuter and hence get more likes. So my best guess would have been this order: pupper > puppo > doggo (actually: puppo > doggo > pupper).

The last question that I addressed was whether there is relationship between the WeRateDogs Ratings and the amount of likes. By calculating the correlation of both variables I firstly got the impression that there is no correlation since the Pearson correlation coefficient was near to 0. After getting rid of the extreme values there is a clear relationship between those variables. With a Pearson correlation coefficient of 49% it is not a very strong correlation but it's ok and it is also visible. In conclusion, this means that either the WeRateDogs Rating is quite a good indicator of the cuteness of a picture or a lot of people like those pictures based on the rating.