Act Report on Twitter dataset Project

Done by Eliel Godsent as a requirement for the Udacity data analysis Nanodegree

This Report highlights some of the findings from my data wrangling project. To begin with, it is important to note that as with every other data set a data analyst would face, the data sets fro this project did not come clean and were thus an intensive one to clean

Three datasets were given:
1. A twitter archive
2. An image prediction
3. Additional twitter data

Twitter archive

This is the beginning dataset that has data on dog ratings, though previously cleaned, still required more cleaning.

Findings:
- Some ratings have multiple dog stages (this is displayed as multiple in the dog-stage column)
- The final cleaned dataset is ~10% of the original twitter archive data
- This data had a tidiness issue and 8 documented quality issues

Image prediction

This file is a prediction score for dog images, it is the result of an already built and pre trained neural net image classier

Findings:
- It wasn't clear of the classification accounted for tweets with two images
- The classifier predicted wrongly on some images

Additional Twitter data

This last dataset was supplementary to the twitter archive to provide retweets and likes

Findings:
- This data has one documented quality issue
- Only 3 columns were necessary in this data (likes, retweets, and id)

PS: the id field was used to join the tables together

Masters table

This is the final cleaned dataset with shape, (218, 10)

Findings:
- A golden retriever is the most popular rated dog breed
- The most liked tweet has 145,055 likes
- The most retweeted tweet has 70,855 retweets
- Pupper is the most common dog stage
- Dog ratings with multiple stages are, 10 for doggo, pupper. And one for doggo, puppy and doggo flooder

Chart is seen here: