# 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 of the dog names were hidden in the text field
- Some of the dog stages were also hidden in the text field
- Some dog names weren't mentioned
- Some tweets had two dog images
- It was necessary to have the dog stages transposed to columns
- Some columns weren't necessary
- Some tweets were' original tweets

**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

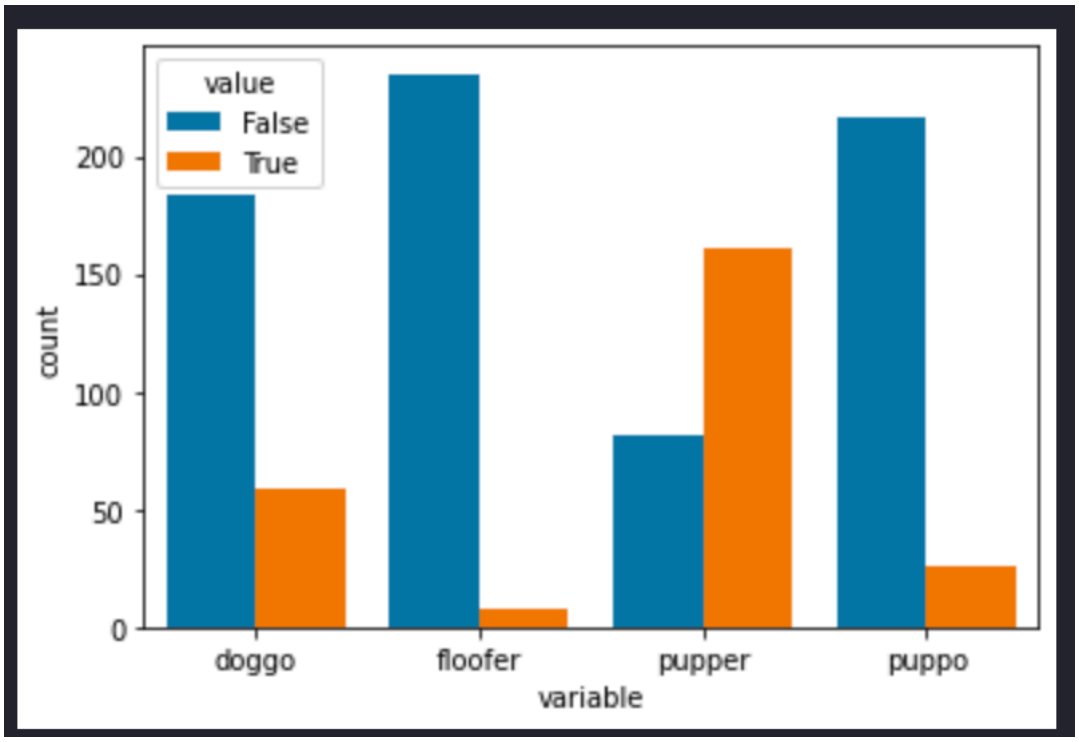
Findings:
- It wasn't clear of the classification accounted for tweets with two images
- The classifier predicted wrongly on some images

**Additional Twitter data**
This last dataset was supplementary to the twitter archive to provide retweets and likes

Findings:
- Some columns are nested dictionaries
- Some columns were not necessary
- The data had tidy issues

General findings:
- Floofers are the least rated dogs with peppers been the most
- A golden retriever is the most popular rated dog breed
- Highest retweets are 79515 and Highest likes are, 132810

Chart is seen below:


value
False
True
200
150
100
50
0
count
doggo
floofer
pupper
puppo
variable