# Wrangle Report

Lucas Belpaire

February 22, 2019

# 1 Gathering data

Gathering the data was done in three large steps.

In the first step the twitter archive was manually downloaded and loaded in using the read_csv method of pandas.

Then, using the Requests library, the image_predictions.tsv file was downloaded programmatically. The imported thing to note here, that contrary to .csv files, all fields were seperated by a tab. This was easily solved by setting the 'sep' parameter of the read_csv method to '\t'.

Finally the extra data (retweet count, favorite count) was collected using the Tweepy library, which returned JSON data. From each JSON object the data that was needed was extracted. This data was then read into a text file. The text file thus contained a JSON object on each line. When all the data was read into this text file, it was loaded in and converted into a dataframe.

# 2 Assessing Data

The assessing process was fairly straightforward. The data was first assessed visually, during this phase some tidiness and quality issues where found. But when assessing the data programmatically most issues where found. The methods most used during the second phase were .info, .describe and .value_counts.

# 3 Cleaning Data

During the cleaning process the quality issues were addressed first. Most issues were solved fairly easily. These included removing incomplete rows, dropping unnecessary columns and changing datatypes. One specific value of a denominator had to be changed aswell. The new value was extracted from the text column of that row.

Fixing the tidiness issues was more difficult. The most difficult issue to solve was adding the dogs_stage column. This was done by combing the 'doggo', 'floofer', 'pupper', 'puppo'. After merging these columns there were to many commas in many cases. Using regex all the columns were cleaned. By using this method it was possible to include multiple stages in the dogs_stage column.