Two data sets, twitter_archive_enchanced.csv and image_prediction.csv have been provided and the third one with additional information from Tweepy had to be obtained.

The first step was to examine the given datasets and create the third one.

Twitter archive dataset had a lot of information, however not all the entries fulfilled the criteria to be original tweets, so those needed to be removed. Few column containing information which wasn't used for analyzing were removed as well. Format of data needed to be fixed.

Archive dataset also had a tidiness issue as dog stage was split to few columns while is should be just one. This part was also fixed with a simple algorithm.

The image prediction dataset had 3 predictions for each tweet and probabilities that the prediction is right. The first True prediction was chosen as a valid one and if none of three predictions were a dog breed, the NaN value would be given. The result of cleaning this dataset was a dataset which only contained tweet_id and predicted breed.

For the additional information from Tweepy, amount of retweets and likes were collected as well as information about the tweet having an image as well as tweet being retweet to eliminate entries which wouldn't be original tweets. After eliminating the entries which either didn't have picture or were retweets, those columns were dropped to result a dataset with information about tweet id, likes and retweets only.

The second step was to unite all three datasets in one. As the only common information for all of them was the tweet id, this column had to have the same format of data. As any mathematical operation doesn't make sense for id, it was turned from being integer to string type.

Result of the union was a table with 13 columns. For better readability the order of the columns which was changed while merging was arranged in a way that the most relevant information as tweet id, name and breed of the dog would come first. For the anaylizing step the new column 'rating' was added which was a result of division of numerator and denominator. Few rating with irrelevant values (based on We rate dogs highest given rating) were either replaced or deleted.