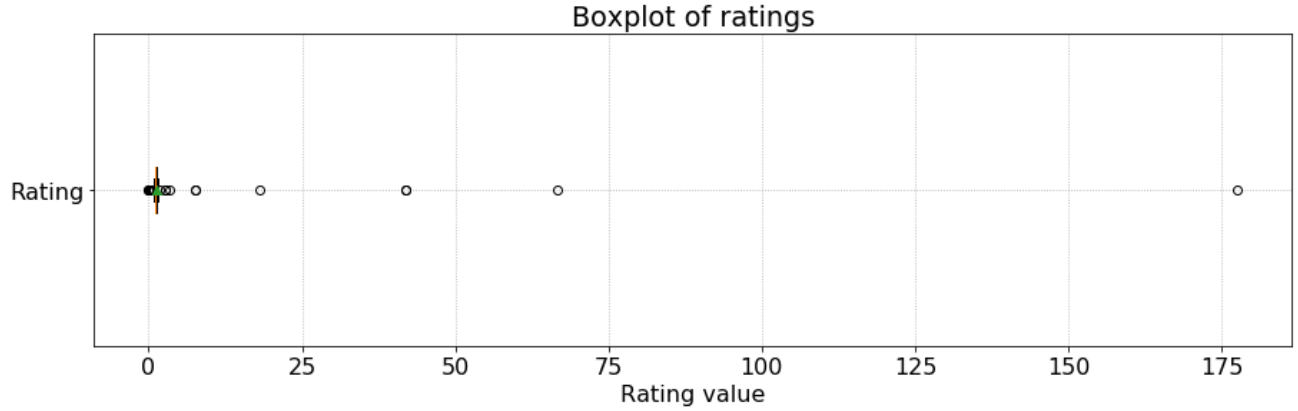
Storing, Analyzing and Visualizing

Insight #1: What is the distribution of ratings? Which dog scored the highest rating?

The boxplot below presents the distribution of ratings after calculating the rating as the ratio: $rating = rating\_numerator / rating\_denominator$.
The average value is: 1.2 and the maximum: 177.6



The dog named Atticus shows the highest rating of: 177.6

Insight #2: What are the types of dogs with highest ratings?

The type of dog is extracted from the predictions p1. We select only the types that appear more than 5 times to make sure to have real dog's breeds. We group the dataset by dog's breed and sort by average rating for a breed only when the confidence of forecast of the breed (p1_conf) is higher than 0.5.

Dog's breed	Average rating
Siberian husky	1.22
Miniature poodle	1.2
Norfolk terrier	1.2
basset	1.2
golden retriever	1.186535
Old English sheepdog	1.183333
Norwegian elkhound	1.175
Eskimo dog	1.171429
tennis ball	1.16
Cardigan	1.16

The Siberian husky is the dogs' breed showing the highest rating with an average of 1.22 for this group. It is followed closely by the miniature poodle, Norfolk terrier and basset with an average rating of 1.2

Insight #3: Which tweet has shown the highest number of re-tweets? And highest number of favorites?

The tweet with the highest number of retweet is the following:
"Here's a doggo realizing you can stand in a pool. 13/10 enlightened af (vid by Tina Conrad) https://t.co/7wE9LTEXC4". It shows a video of a dog swimming in a pool.
The tweet with the highest number of favorite is the following:
"Here's a super supportive puppo participating in the Toronto #WomensMarch today. 13/10 https://t.co/nTz3FtorBc"

Conclusions

This real dataset coming from tweet data can be very messy and dirty. A lot of work is required to tidy and clean the dataset before being able to extract any valuable information from it. The project helps realizing the importance and amount of work required to clean and tidy a dataset based on non-simulated data.
The insights showed that Atticus shows the highest rating of: 177.6 with its handsome costume. The Siberian husky is the dogs' breed showing the highest rating with an average of 1.22. And the tweet showing the highest number of retweets shows a video of a dog realizing that you can stand in a pool. I also realized that the data contains a significant number of tweets not related to dogs with the lowest ratings being a hen or an electric fan...