Visuals and insights over weratedogs

We all know the importance of data in our modern life and how data affects all what is going around us. That is why we always need our data in its best form, so we can get our knowledge and make predictions for the upcoming events using this data.

However, what happens if we get our data messy or having default entries or some quality issues, etc.?

That is the point where we have to stop and start analyzing and wrangling our data to get the most out of it, today we have an example with showing a lot of gratitude for Twitter account weratedogs for their support with the following material

Working on such data made us look how we can get the best out of what we have.

Before starting, we had some issues in our data so we decided to assess our data and found the following:

- Wrong values shown while extracting ratings
- Float values in some ratings
- Some tweets had combined dog ratings
- Dog stages was including a None values
- Some dog names were extracted by the wrong means and some had a none value
- Some entries had no images that we had to drop
- false predictions for some of the images

We started our wrangling process and did the following:

- Extracted the right values
- Changed the column type to float
- Divided the rating over the number of dogs
- Replaced all none values with NaN
- Extracted the dog names and replaced post with no name with NaN
- Removed all entries with no images
- Removed all false predictions

And we got an expectacular results let's see some of these results
Before cleaning dog stage values we have 1971 values 1668 `None` value but afterwards we were capable of categorizing 303 values into specified dog stages and shown in the following charts.

We started to ask what the most posted dog stage is.

We found the pupper dog stage is the most common stage of all

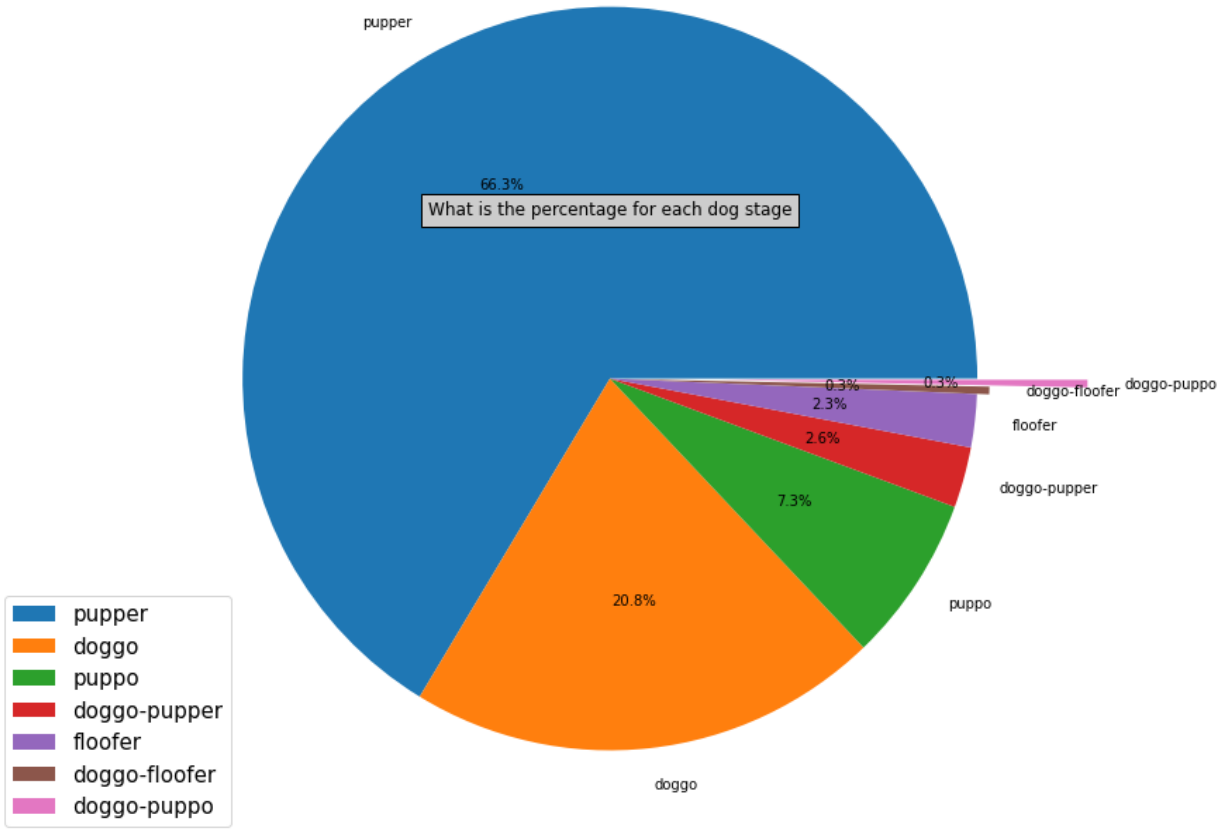

Figure 1: Dog stage percentage

What about the favorite dog stage?

We found from the data that the most retweeted and favorite dog stage is the doggo-pupper

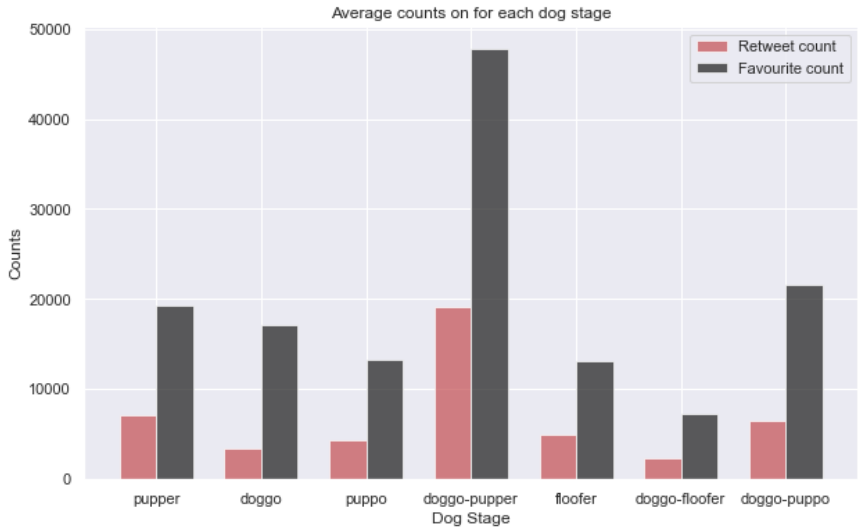

Figure 2: average counts per dog stage

And then we started investigating forward using the dogs breed predictions and found an amazing thing the most retweeted breed was totally different from the favorite one

The most retweeted breed is the Standard poodle

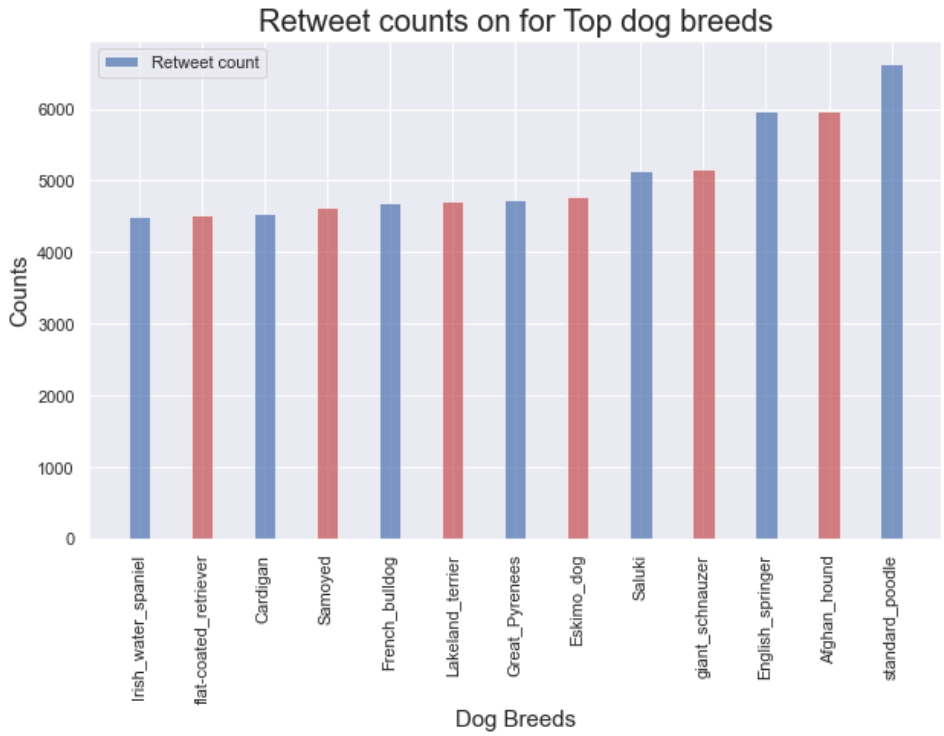

Figure 3: Most retweeted dog breeds

Meanwhile the favorite is Saluki

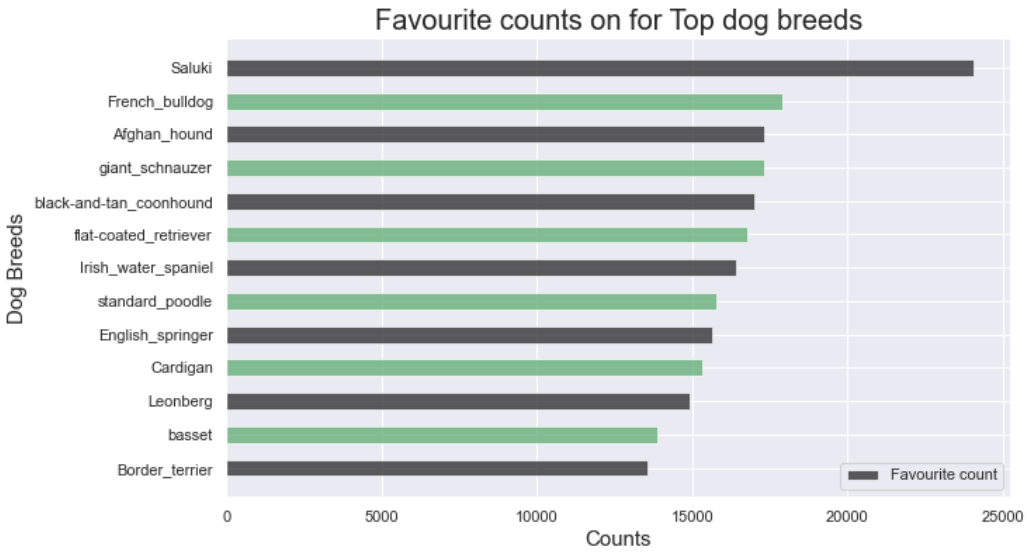

Figure 4: Most favorite dog breed

At that point we started to have a question 'how precise is our data after the cleaning process' so we had to dig deeper and to go for the extra mile in our analysis

What we found was very interesting and made our data more valuable and validated it, let's have a look over the tweets that have dogs pictures in comparison with the ones with no picture and the result was astonishing, number of retweets and favorite count was higher for the ones with dog picture .

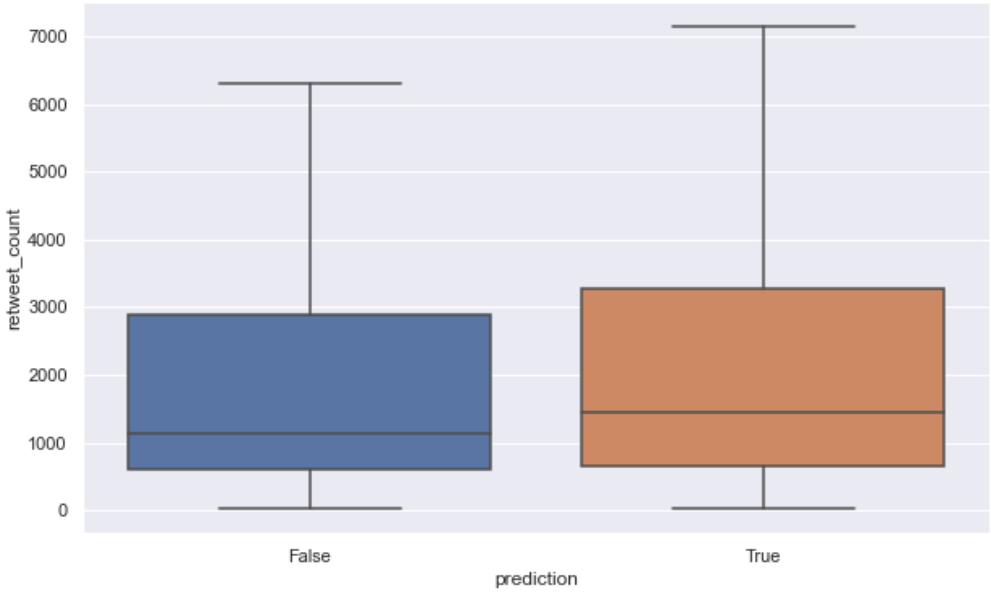

Figure 5: retweet count for tweets with and without dog picture

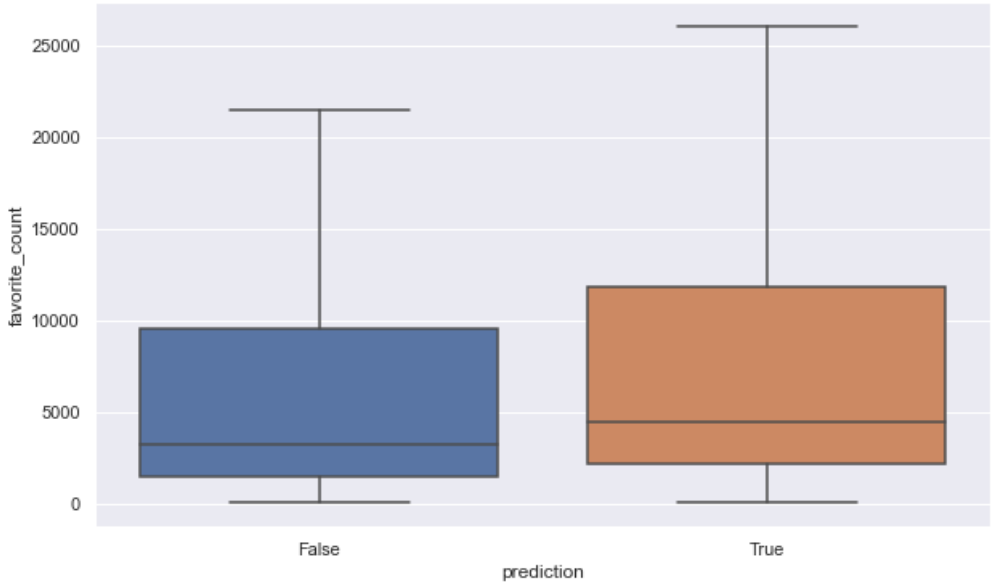

Figure 6: favorite count for tweets with and without dog picture

And at that point we validated our clean data and we saw the results of the cleaning process