

Data Wrangling Report

This is the `act_report` it contains all the insights and displays the visualization(s) produced from my wrangled data.

Observations For Enhanced Twitter Archive

Quality

- * There are retweets present in the data
- * Dog names are not consistent
- * ID variables are sometimes integers or floats (numeric)
- * Column names are not always meaningful
- * Retweeted_status_timestamp is not a datetime variable
- * Source values are formatted as `<a href=url <a/>`
- * Some rating numerators less than 10
- * "retweeted_status" variables are numeric
- * Some rating denominators are not equal 10

Structure

- * Doggo, floofer, pupper and puppo all refer to the same measurement "dog stage"
- * Source and expanded urls have several information
- * There is more than one stage filled for individual dogs

Observations For Tweet Image Predictions

Quality

- * Some column names contain '_' and '-' instead of spaces
- * Some names start with an uppercase while some start with a lowercase

Observations For Twitter API

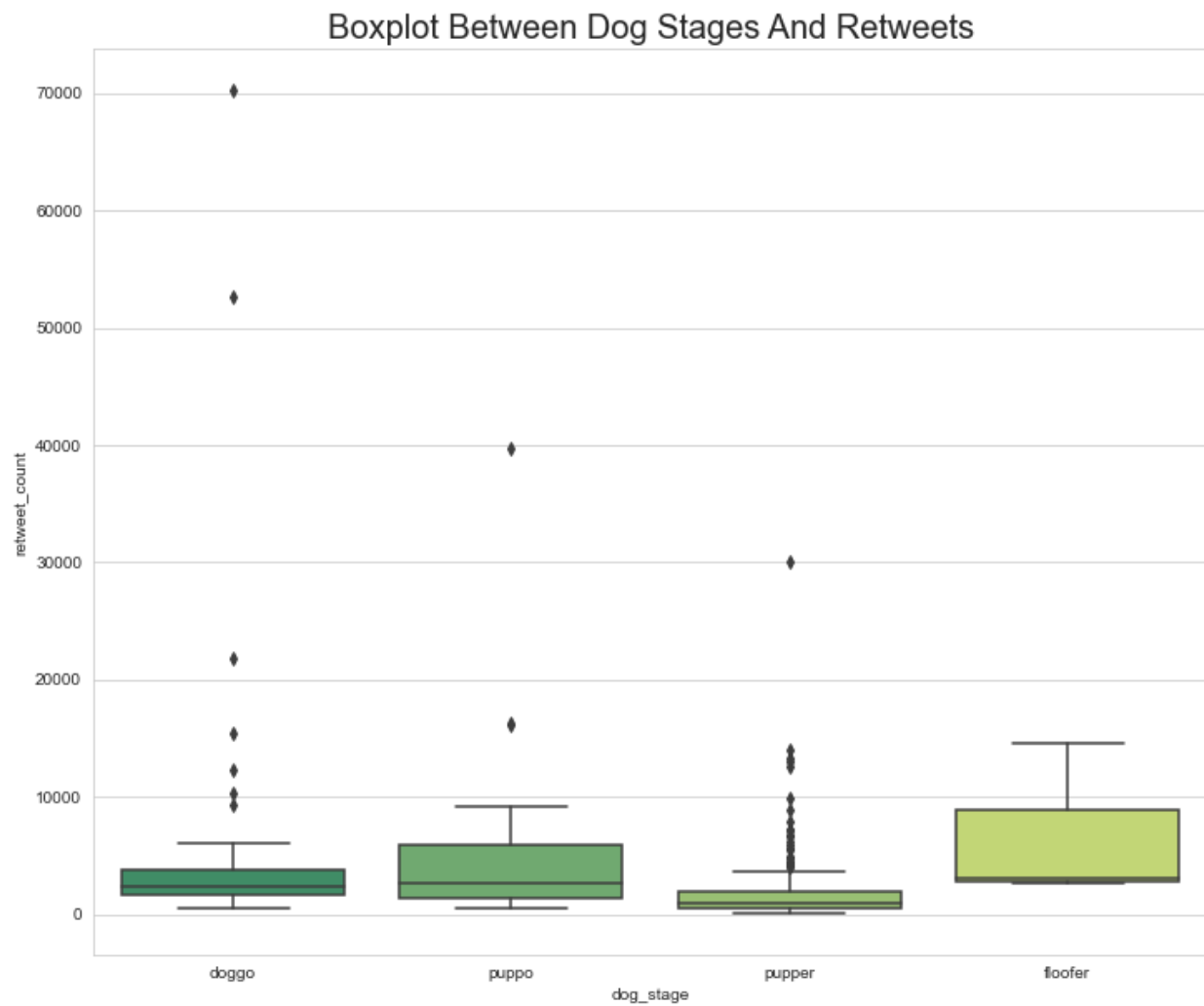
Quality

- * Missing data (the archive dataset has 2356 ids but only 2354 show up)

I. Relationship Between Favourite_count and Retweet_count

Analysis

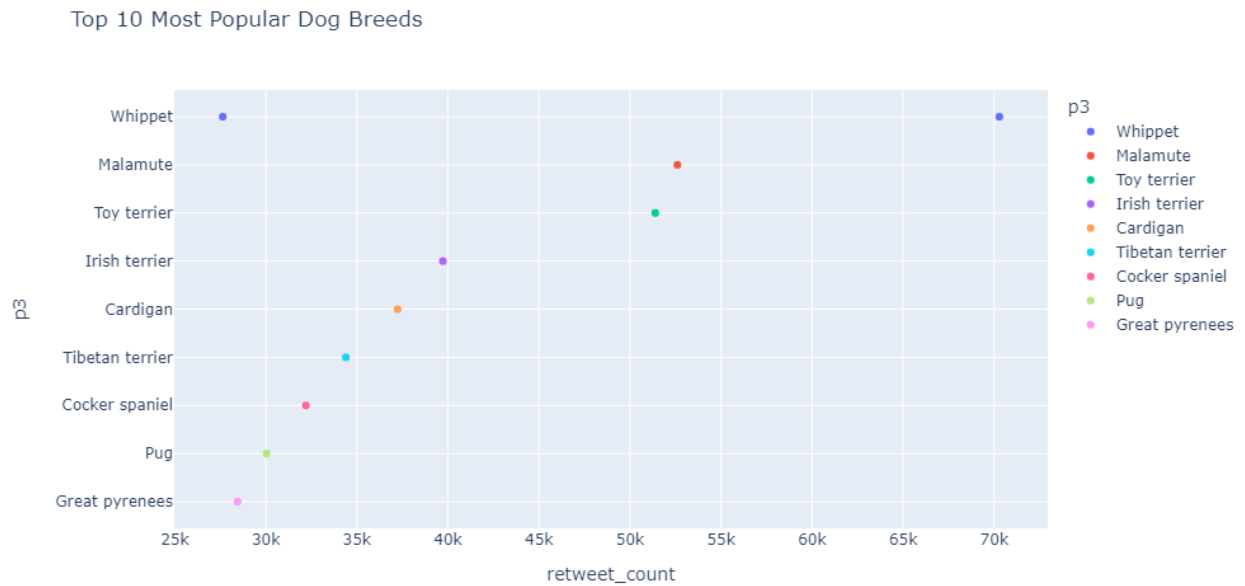
- There are more dogs in puppo category
- There are more retweets in doggo category



II .Top 10 Most Popular Dog Breeds

Analysis

- Whippet is the most retweeted dog breed



III . Amount of Each Dog Stage

Analysis

- Pupper has the highest percentage
- Floofer has the least percentage

