Act Report

Lucas Belpaire

February 22, 2019

1   Main Questions

Three questions were briefly examined.

1. Which species scores the highest points on average?

2. Does the favorite_count correlate with the score?

3. Does the retweet_count correlate with the score?

2   Which species scores the highest points on average?

We examined this statement visually. The 'p1' column is used to know which species each entry is. p1 is chosen because it is the most accurate guess according to the algorithm. All rows that the don't have dog species in this column were dropped.

The species were grouped together and a mean score was calculated for each group. This data was then plotted using a bar chart (see figure 1). There are two outliers (clumber and japanese spaniel). The fact that they deviate so much from the average is probably caused by the fact that for both species there is only one sample in the dataset. Their averages are thus not representable for the entire species. We can also clearly see that there is very little variance in the mean score between different species.
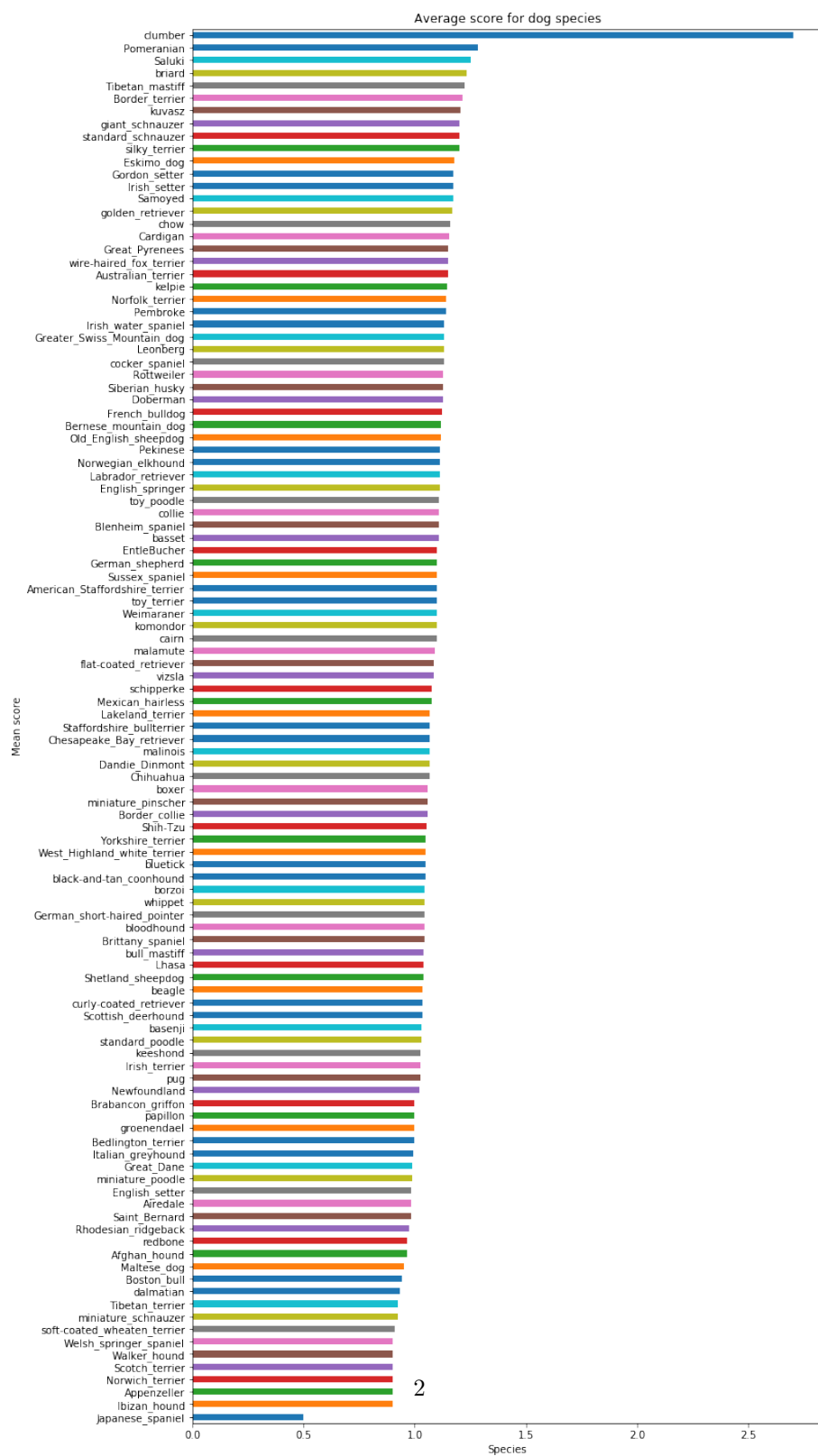
Figure 1: Average score for dog species

3   Does the favorite_count correlate with the score?

This question can be answered visually and by calculating the Pearson correlation. The best way to visually represent two variables is by plotting a scatter plot. This can be seen in figure 2. When looking at the figure no clear correlation can be seen.
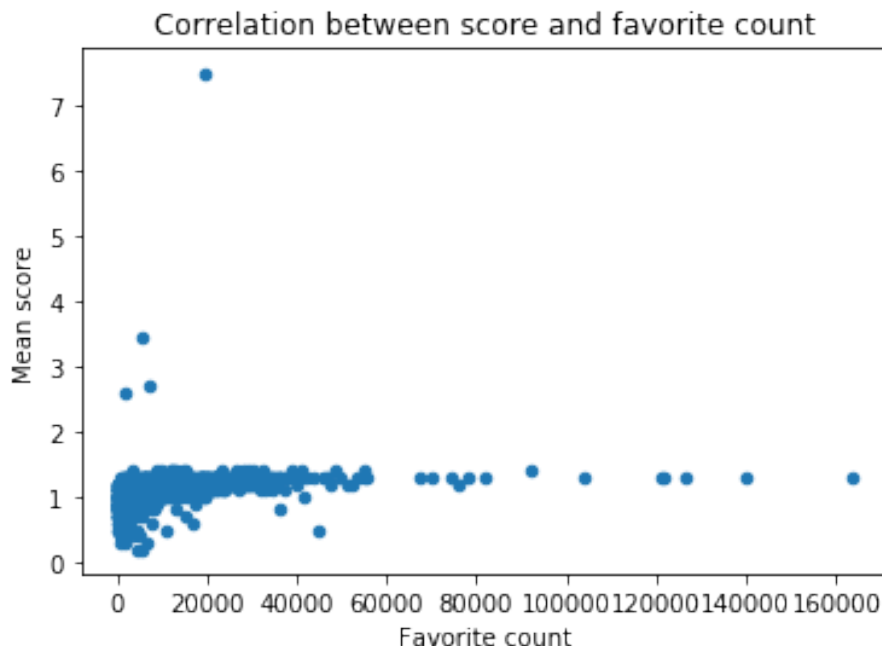
Figure 2: Correlation between score and favorite count

The Pearson correlation coefficient has a value of 0.286, which is low and thus indicates no correlation.

4   Does the retweet_count correlate with the score?

This question is fairly analog to the previous one, thus an analog solution is used. The scatter plot and Pearson correlation coefficient, which has a value of 0.220, indicate that there is no correlation between retweet count and score.
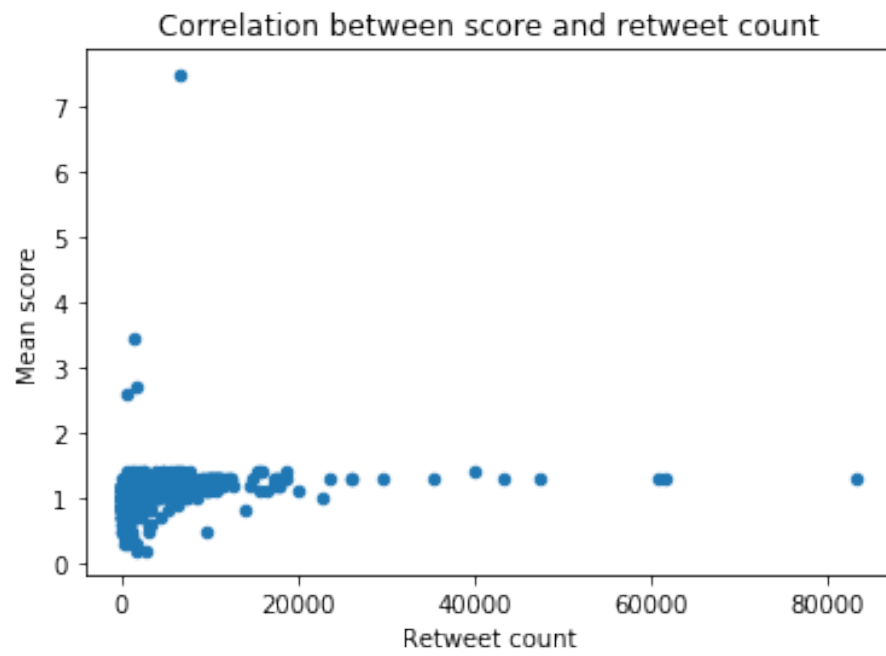
Figure 3: Correlation between score and retweet count