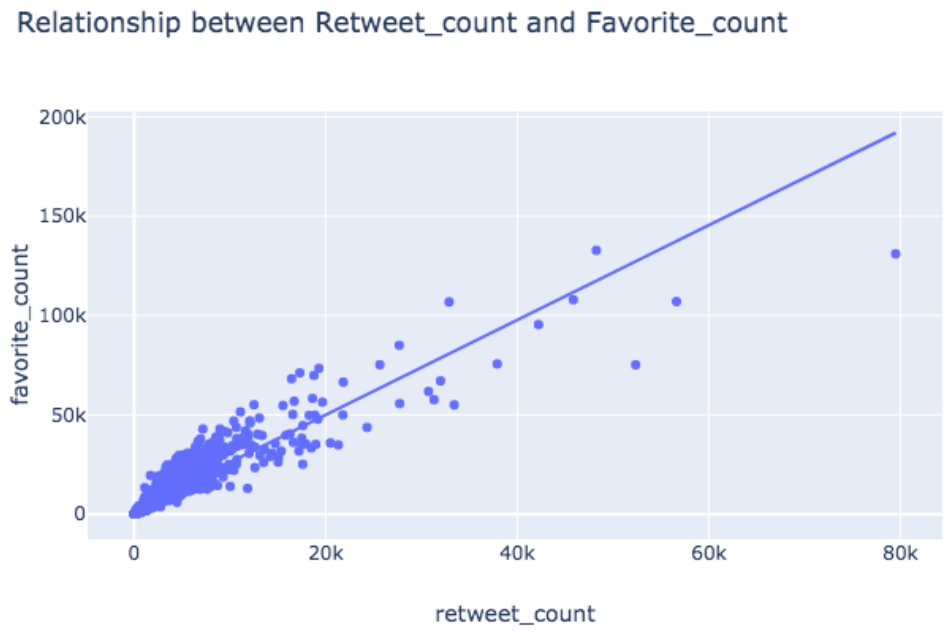
Act Report

According to the combined data set of 'archive', 'prediction' and 'twitter_json', five questions were posed, which are as follows:

1.the relationship between retweet_count and favorite_count

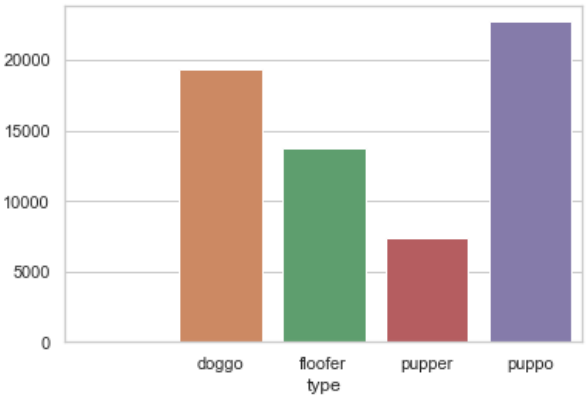
The solution is to calculate the correlation efficient for retweet_count and favorite_count and then plot a scatter for them.

The correlation efficient is higher than 0.7, which means retweet_count & favorite_count have a stong uphill linear relationship

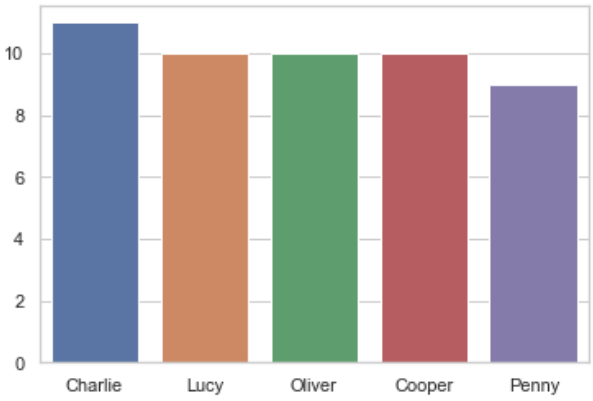


2. which dog type has the highest mean favorite_count

Puppo has the highest mean favorite_count of around 22724, pupper has the lowest mean favorite_count of around 7424, it shows that dog type really matters when it comes to the favorite of users.



3. the top 5 common dog names

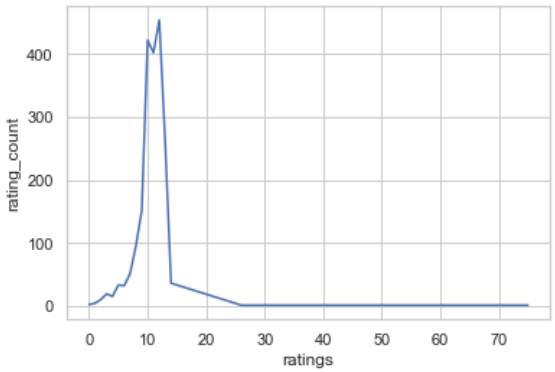


It can be seen from that Charlie, Lucy, Oliver, Cooper and Penny are the top 5 common dog names, according to the data set, Charlie has the highest count of 11, Lucy, Oliver and Cooper come after it with a count of 10, Penny has the count of 9.

4. comparison of the correct rates against dog of p1, p2, p3

p1 has the highest confidence towards dog type, however, the correct rate of which is slightly lower than that of p2. Therefore, the confidence towards dog type and correct rate dont have a strong relationship.

5. presentation of the distribution of ratings



Above is the line chart with two highest rating removed from the dataframe, it can be seen from that ratings around 10 have the higher count. According to the dataframe of ratings, ratings of 12 has the highest count, and the ratings of 10 & 11 come after it.