Analysis

-Since we dealing with dataset with 1968 rows × 23 columns
-Our data set about tweets related to the dog images and rationing them. We have a dog_bread column(doggo, floofer , pupper, puppo) and we have 3 classifiers that predict the type of the dog in the image .
-now we will try to ask some questions and try to find their answers

1- since we have the dog_bread column we could use it to know which bread has the highest retweets and favorite count .
2-we could use scatter plot to know which features are positively correlated with retweet count and favourite count .
3- we could use prediction that has the highest performance .
4- the classifier that has the highest performance , we could use it to know which type of dog gets the highest retweet and favorite count .
5- how could followers affect raring ,retweets and favourite_count.

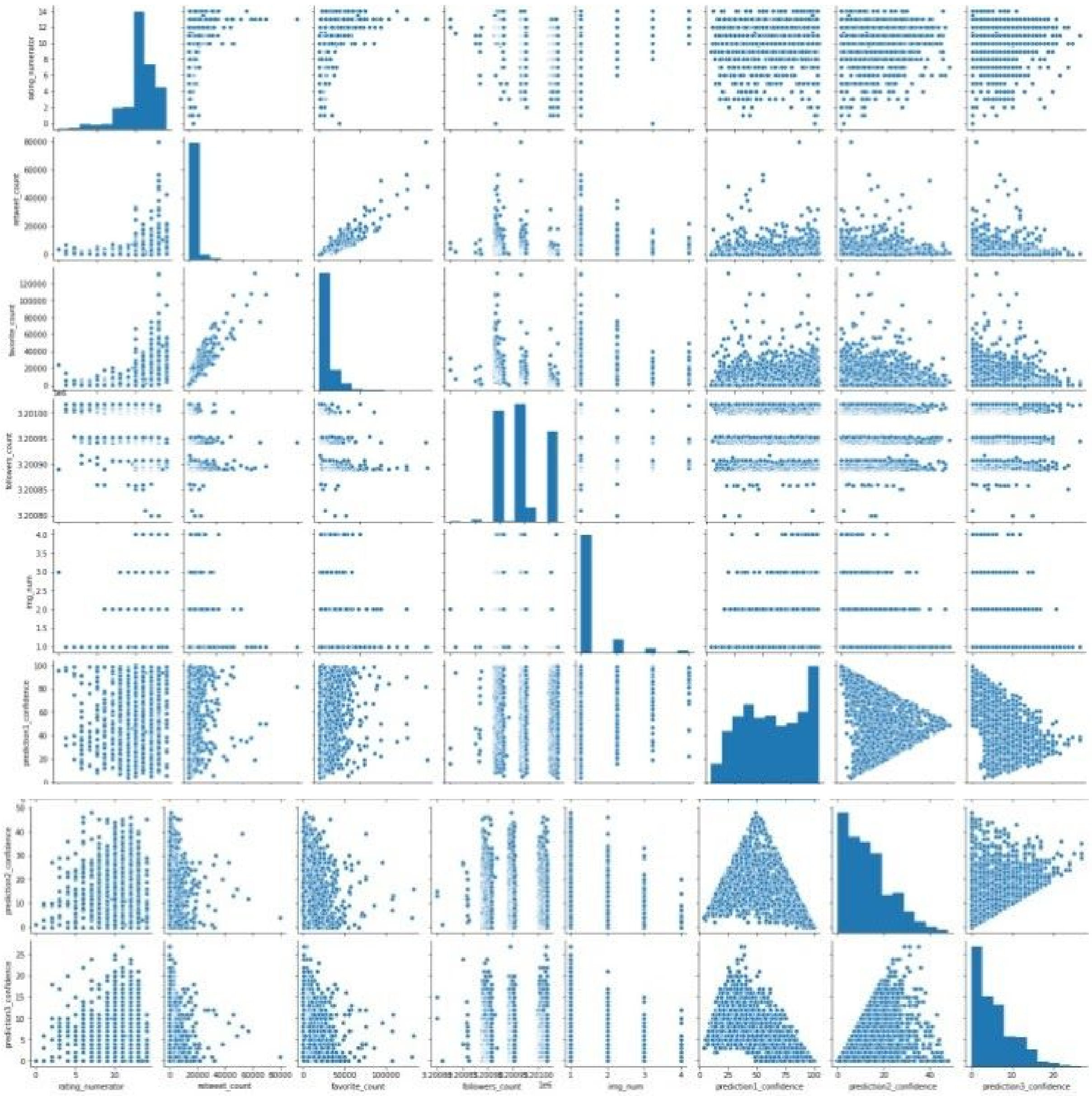
1-
We use the pd.corr() method and scatter matrix of seaborn to get the correlation between retweet_count feature and the other numerical features and between favourite_count and the other numerical features .
We notice that:
Retweet_count has:

- Positive strong correlation with favourite_count
- weak positive correlation with rating_numoritor
- Weak negative correlation with follower_count
- Weak positive correlation with prediction 1
- Weak negative correlation with prediction 2
- Weak negative correlation with prediction 3

Favourit_count has:

Positive strong correlation with favourite_count

- weak positive correlation with rating_numoritor
- Weak negative correlation with follower_count
- Weak positive correlation with prediction 1
- Weak negative correlation with prediction 2
- Weak negative correlation with prediction 3



2-

Now we will deal with dog_bread:

-We have 7 types of dog_bread in our data frame so we will filter data frames (pd.query)depending on them and use describe() to compare between them.

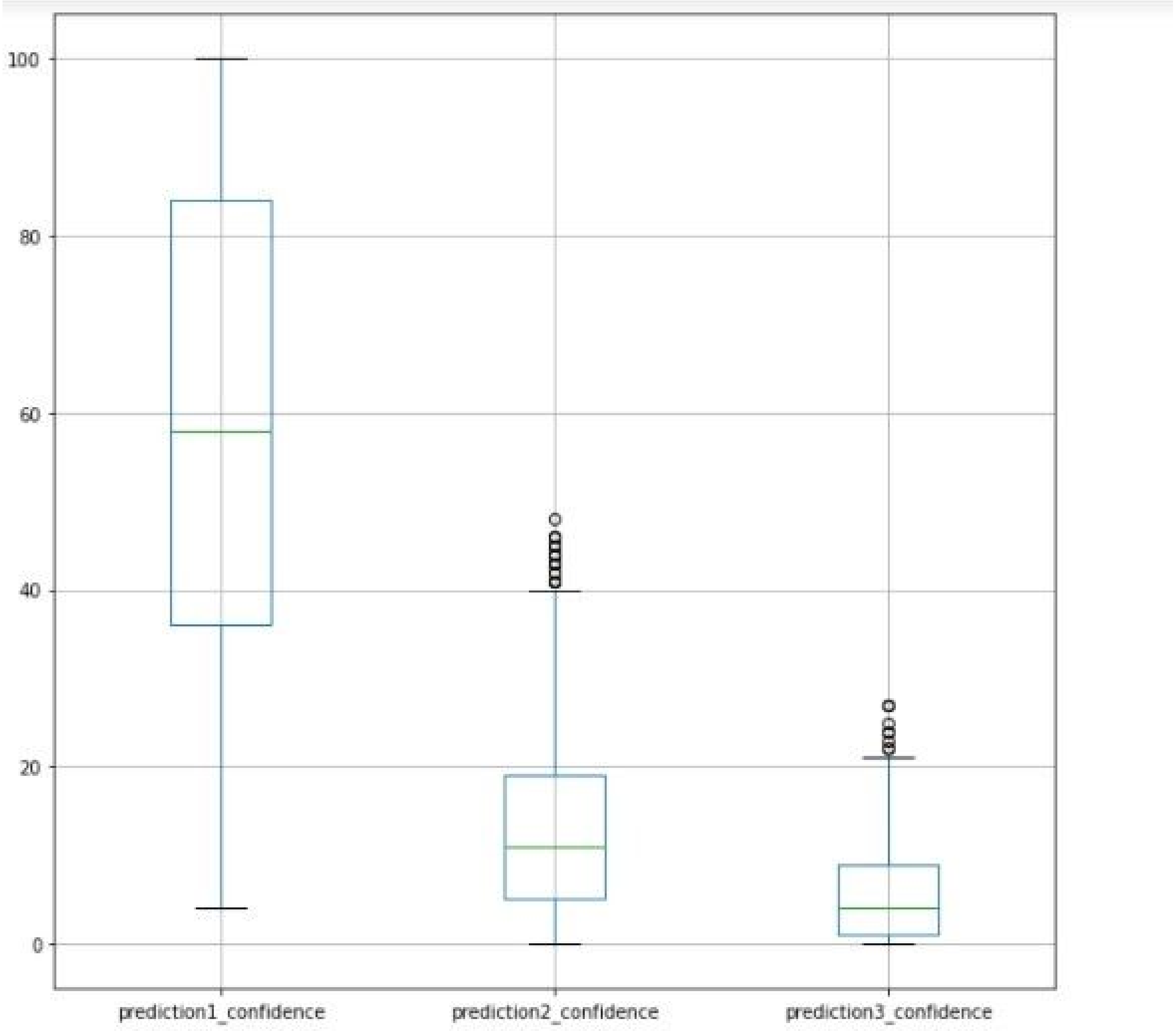
-We notice that:

1- even we have 1 observation of doggo_puppo but we notice that it has the largest value in retweent_count and favorite_count and a very high rate.

2-doggo & puppo have high values and are close to each other.

3-depends on the value of the mean pupper has the lowest values.

3-
Box plot for the tree predictions columns



We notice that it is clear that prediction1 has the highest accuracy than the other two predictions .

4-
Since the prediction1 has the highest accuracy we will use the most 5 frequent types in it using(pd.query) and measure them with rating and retweets and favourite count.

We notice that golden_retriever is the most frequent and it has almost the largest values in both mean and max. And vice versa pug is the least frequent of the 5 types that we choose them and almost it has the min in both mean and max values.

5-
We use follower_count to know how could the followers cloud affect the raring and retweet and favourite so we filter it into 3 categorical depend on the 5 summary statistical number 75%,25%,50%using (describe() and (pd.query)) and we notice that the lower the account followers the higher the rating ,retweets and favourite_counts and vise versa the larger follower the lower rating ,retweets and favourite _counts .