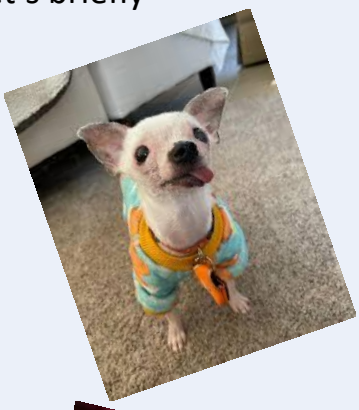
Insights and Visualizations Report

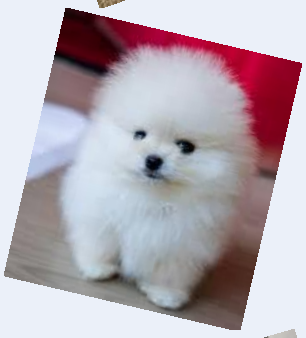
Introduction

After the wrangling phase, we have done some EDA and concluded some insights on our master_df (The merged dataframe), but first let's briefly explain what the data is mainly about.



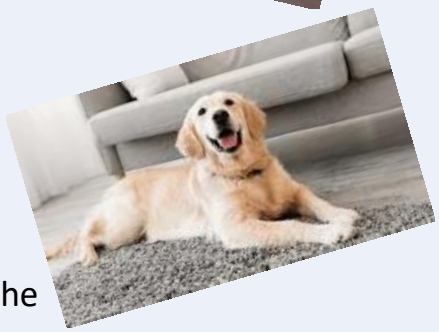
Brief summary about the topic of the data

This data was gathered from X (previously twitter) API from the WeRateDogs page. This page has a lot of participants who post tweets with photos of their dogs, and write a short comment on it, then the followers have to rate the dogs out of 10, as shown in the photos.



The **tweet-json.txt** and the **twitter-archived-enhanced.csv** datasets provided by Udacity have many other info, such as:

1. Info about the tweet (time of post, full text, expanded urls, ….)
2. Ratings of the audience
3. Number of retweets
4. Number of favorites
5. The length of the tweet in chars



And much more information about the tweets

The third dataset **image-prediction.tsv** have the data of result of prediction of three machine learning models of the type of the dog present in the tweet.

What are our analysis questions?

The questions which we aim to answer through our analysis are:

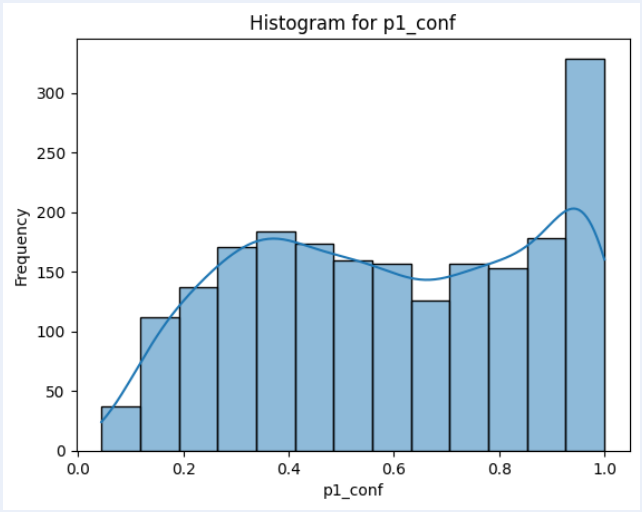
1. Which ML model is the highest in terms of being sure of the dog kind?
2. What is the spread of the ratings?
3. What is the distribution of the favorites?
4. What is the spread of the number of retweets?
5. What is the relation between retweets count & favorites count?

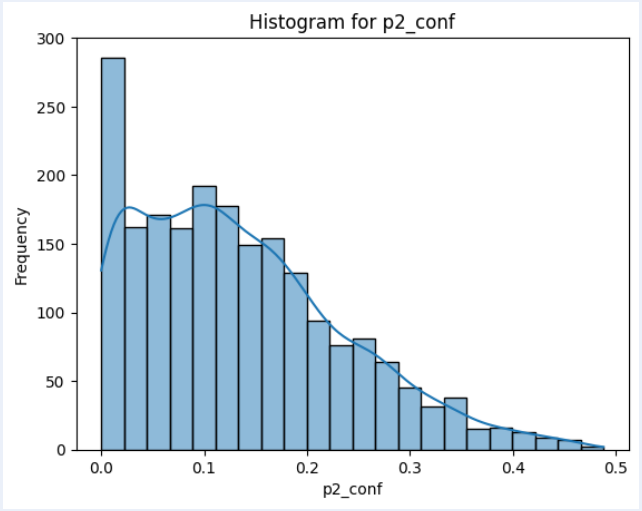
6. What is the relation between ratings and favorite count?

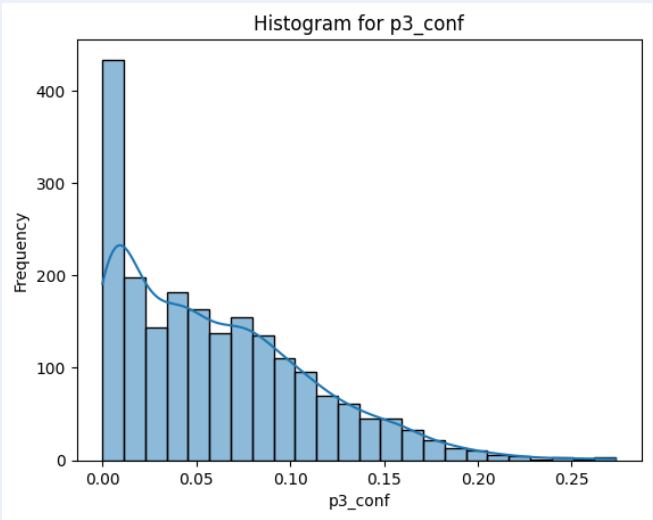
Research question #1: Which ML model is the highest in terms of being sure of the dog kind?

Expected: I expected that the 3 models will be close in results to each other

Visualization(s):





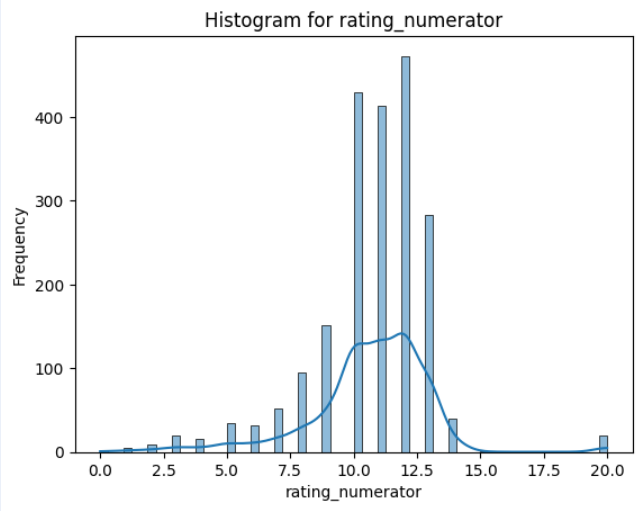


Insights: The p1 model is the highest as its results are very close to 1, while p3 model is the least model as it results didn't even pass 0.25

Research question #2: What is the spread of the ratings?

Expected: I expect that most ratings will be around the mean

Visualization(s):



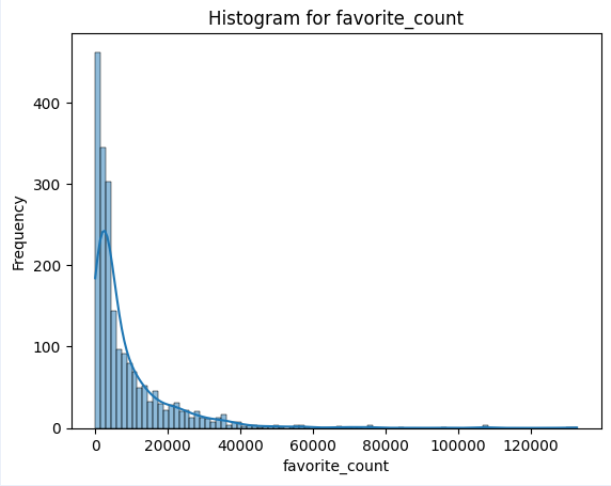
Insights: As expected, the ratings are around the mean which is about 10, so this data is symmetric

---

Research question #3: What is the distribution of the favorites?

Expected: I expect that the favorites count will be around the mean

Visualization(s):



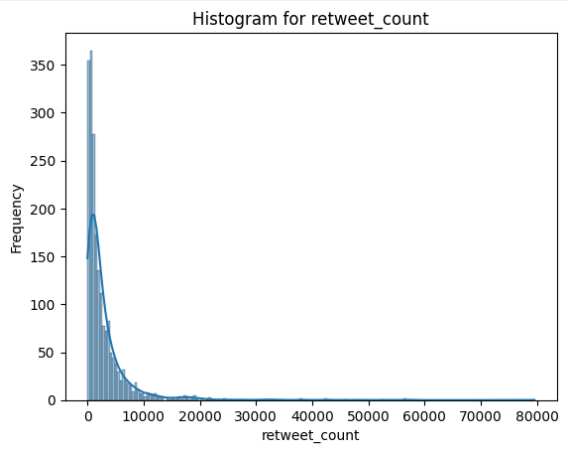
Insights: Unexpectedly, most favorite count are between 0-40000, and there is a lot of outliers that range between 40000-120000 which is a large range. The data is very right skewed

---

Research question #4: What is the spread of the number of retweets?

Expected: I expected that the number of retweets will be around the mean

Visualization(s):



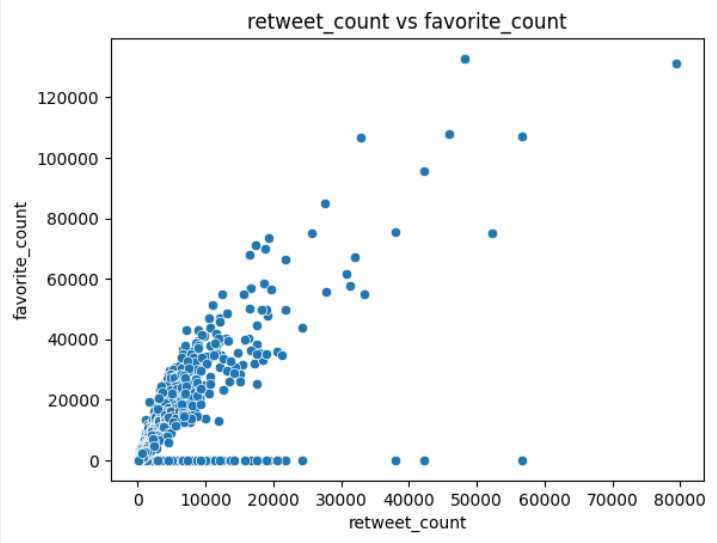
Insights: Unexpectedly, most retweet counts are between 0-20000, and there is a ton of outliers between 20000-80000, which is a large range. This distribution is very right skewed

---

Research question #5: What is the relation between retweets count & favorites count?

Expected: I expected that there is a slight correlation between retweet & favorite counts

Visualization(s):

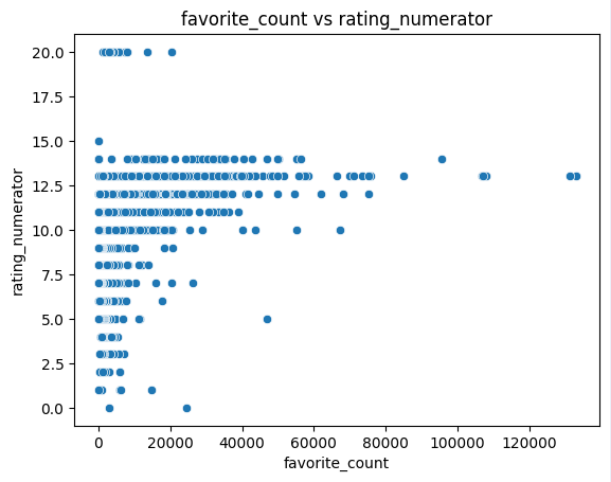


Insights: As expected, there is a slight correlation between both of them

Research question #6: What is the relation between ratings and favorite count?

Expected: I expected that there is a correlation between ratings and favorite count

Visualization(s):



Insights: Unexpectedly, there is a very low correlation between themS

Conclusion

In this document, we discussed:

- A brief summary of the topic of the data
- Our analysis questions that we aim to answer
- Each question with its answer

We can summarize the insights in the following table:

Question	Insight
Preference between ML models	p1 model is the best, p3 model is the worst
Rating distribution	A symmetric distribution
Favorites distribution	A very right skewed distribution
Retweet distribution	A very right skewed distribution
Relation between favorites & retweets	A slight correlation between them
Relation between rating & favorites	Almost no correlation between them