INSIGHTS REPORT

Udacity Data Analysis Course

Remi Dongmo

Introduction

Real-world data rarely come clean. Using Python and its libraries, we gathered data from various sources and formats, assessed its quality and tidiness, then cleaned it.

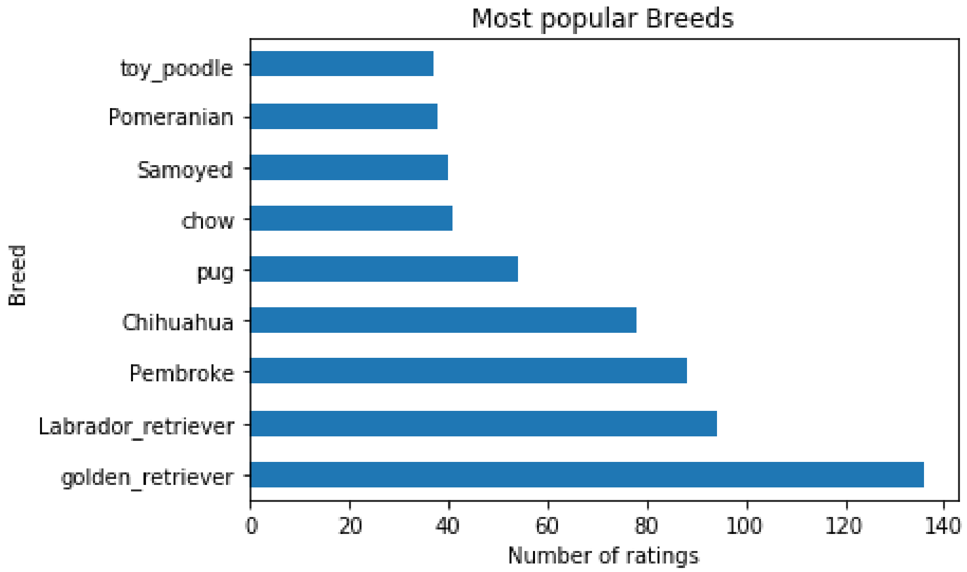
Now we will communicate the insights that we got after our wrangling. We will also show some visualizations made during the analysis process. For every insight we started by a research question, then proceeded with investigation coding and finally drawing out conclusions.

The analysis involved variables obtained from the 3 datasets

1. The most popular/common breeds

We judged the popularity of breeds based on their counts. The most times a breed is rated the most it is popular or likely to find. The chart beside illustrates our analysis:

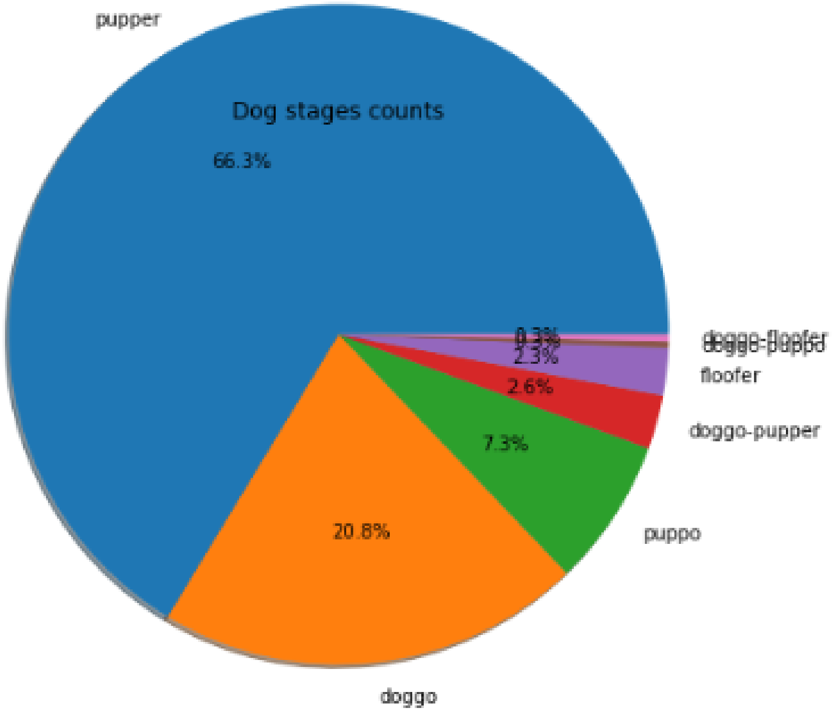
- Golden retrievers are the most common breed
- Followed by Labrador retrievers

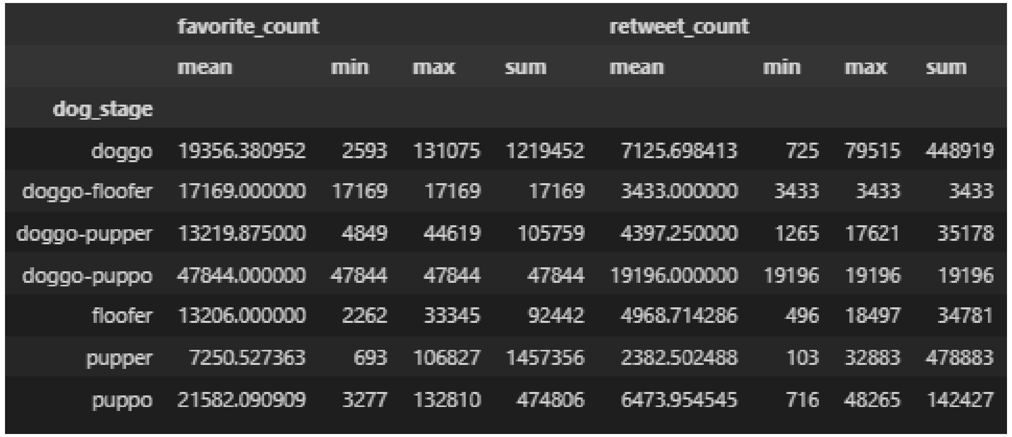


2. The most publicly liked Dog stage

Our next interrogation was about analyzing the most appreciated dog stage. Not necessarly the most popular or common. We would base our analysis on all the users feedback and not the account holder’s ratings. The variables we needed for that were favorite counts and retweet counts. The figures beside shows a summary of relevant metrics. Ignoring the combined dog-stages:

- The majority of ratings are done on puppers but on average they have the less favorite and retweet counts, people probably prefer what is more rare
- Doggos appear to be the most likely accepted dog stage,



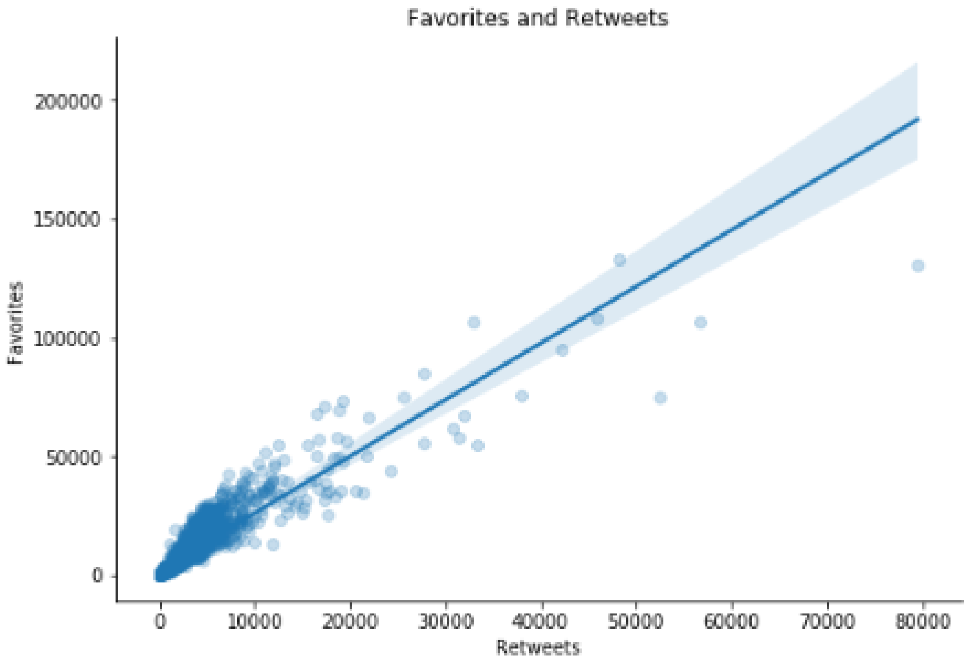


	favorite_count				retweet_count			
	mean	min	max	sum	mean	min	max	sum
dog_stage								
doggo	19356.380952	2593	131075	1219452	7125.698413	725	79515	448919
doggo-floofer	17169.000000	17169	17169	17169	3433.000000	3433	3433	3433
doggo-pupper	13219.875000	4849	44619	105759	4397.250000	1265	17621	35178
doggo-puppo	47844.000000	47844	47844	47844	19196.000000	19196	19196	19196
floofer	13206.000000	2262	33345	92442	4968.714286	496	18497	34781
pupper	7250.527363	693	106827	1457356	2382.502488	103	32883	478883
puppo	21582.090909	3277	132810	474806	6473.954545	716	48265	142427

3. Are favorite counts correlated to retweet counts?

Next we wanted to understand if being more retweeted meant a rating was also more liked (or added to favorites), the variables necessary for this was retweet_count and favorite_count. We plotted a graph to visualize how they were comparing, and we could conclude that the:

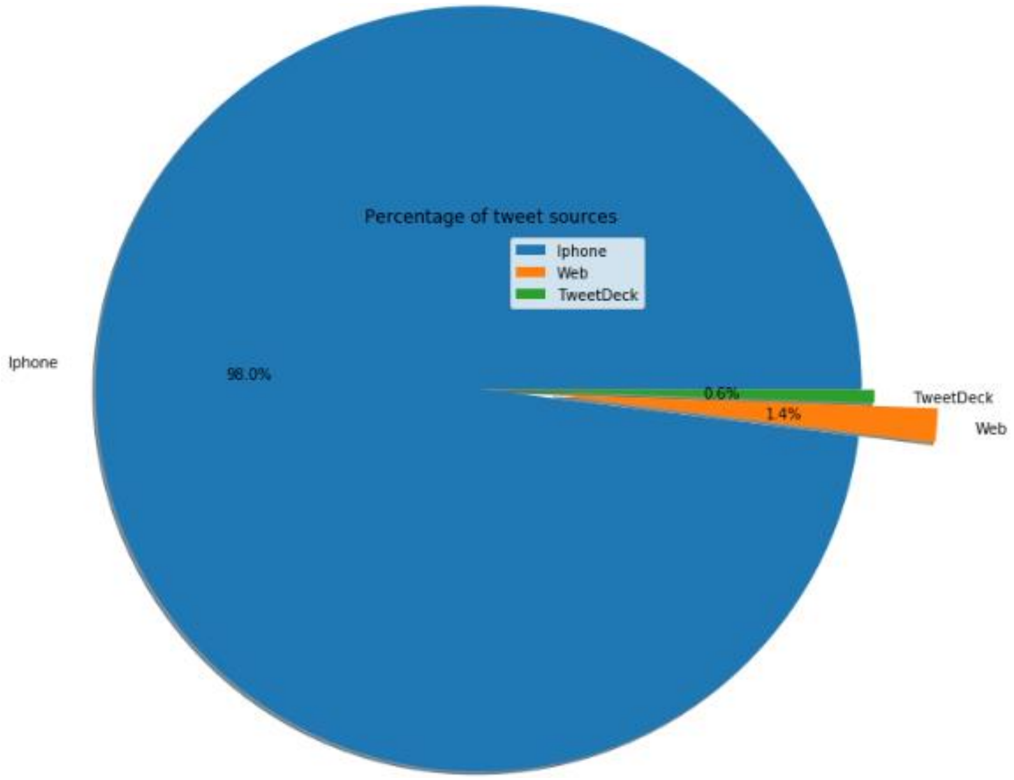
- The correlation between favorite counts and retweet counts is positive, hence they relate.



4. From which software source are the ratings made the most?

Finally we wanted to analyze how often the twitter account was using his different softwares for ratings. The study was based on the source variable which could be either from Twitter Iphone, twitter web or TweetDeck:

- Almost the totality of the ratings are made from an Iphone



Conclusion

This is the end of our analysis process. We could make many interesting insights and visualizations thanks to the preliminary gathering and cleaning process, more insights could have been made but let’s just keep it brief for the moment.