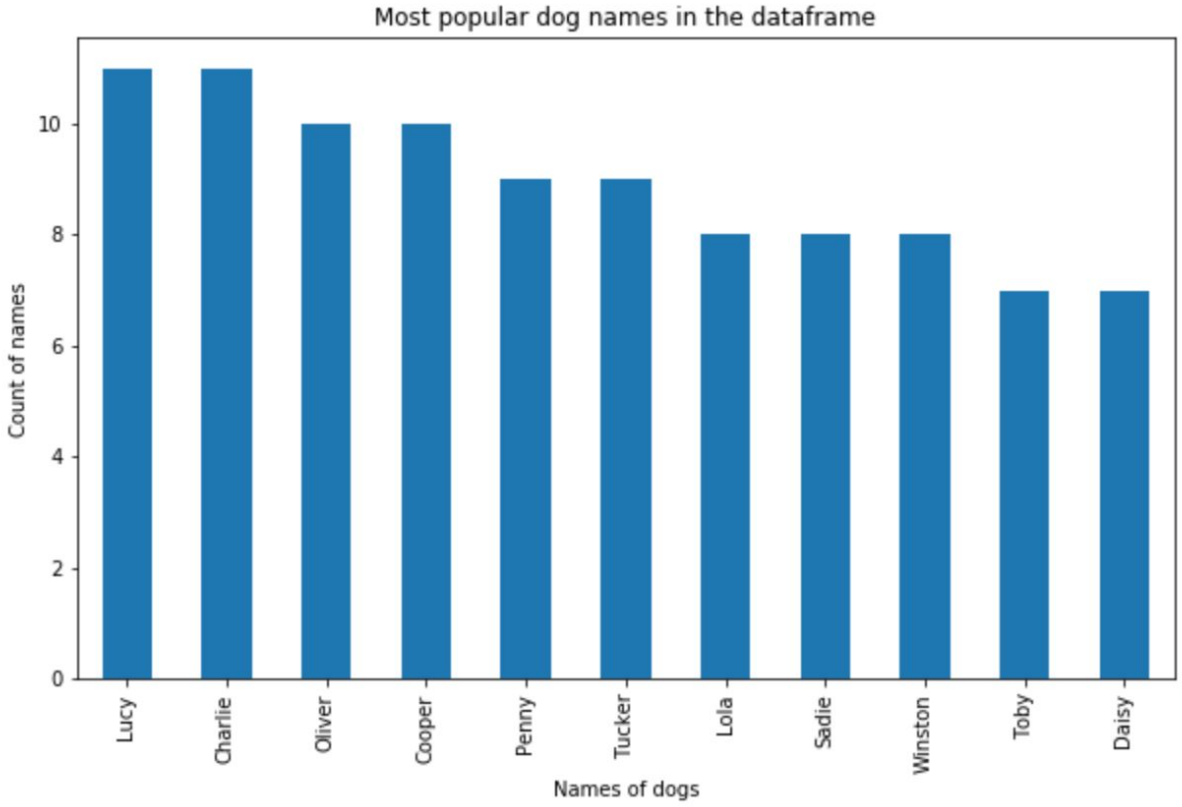
Analysis and Visualisation of the project “WeRateDogs”

After gathering and cleaning datasets from Twitter archive WeRateDogs, I combined them into one final dataset. I will be performing analysis based on the final dataset. I will be performing basic exploratory analysis and for visualisation I will be using matplotlib and seaborn.

I provided 2 visualisations and three insights in my report.

Most popular dog names



753 dog names in the dataset are unknown, so I excluded from from visualisation. Top dogs names include: Lucy, Charlie, Oliver, Cooper, Penny, Tucker, Lola, Sadie, Winston, Toby, Daisy.

Insights on different dog stages

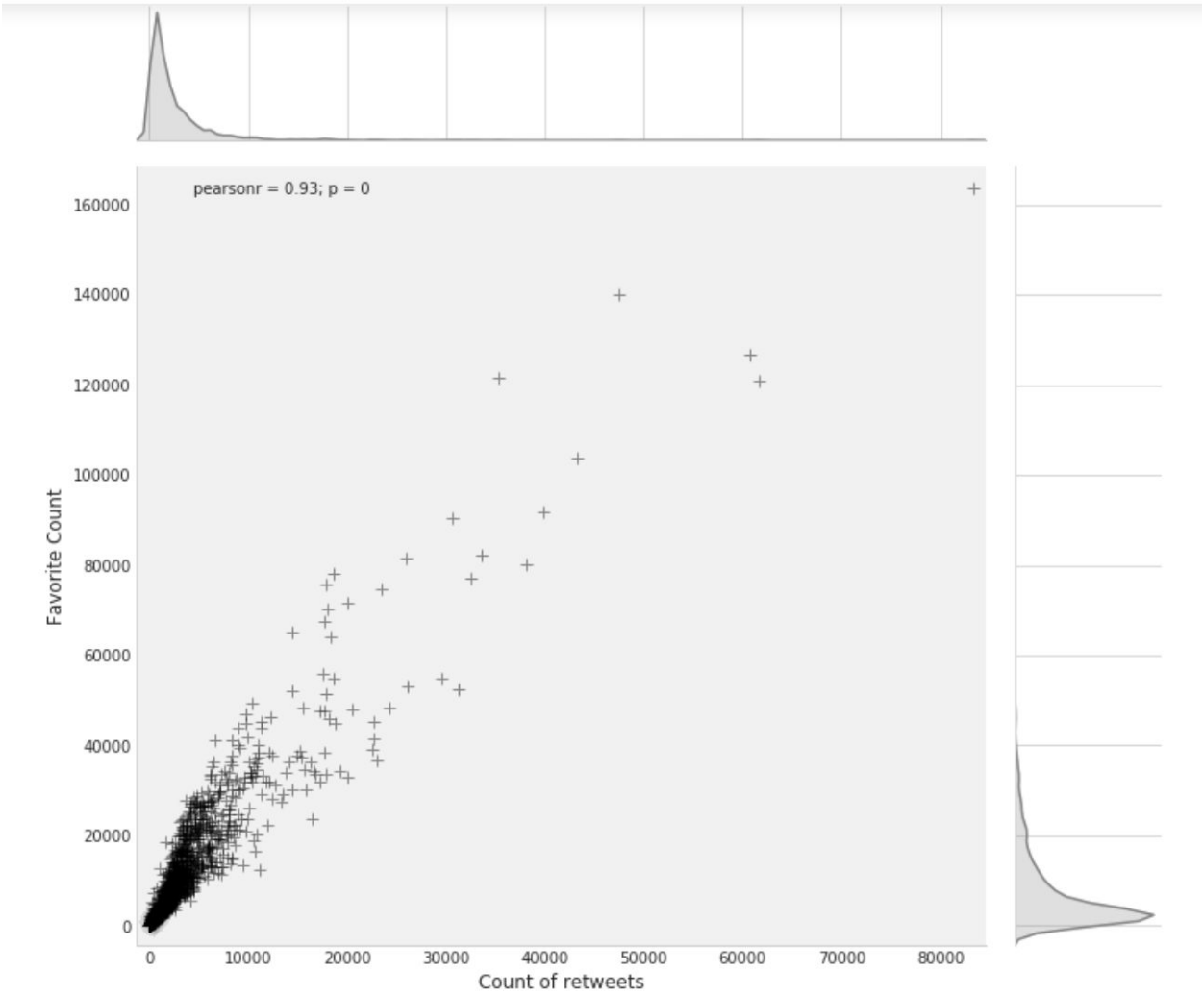
My dog stages includes: doggo, floofer, pupper, puppo, doggofloofer, doggopuppo. I grouped the dataset by stage and compared their rating.

stage	count	mean	std	min	25%	50%	75%	max
None	1831.0	inf	NaN	0.0	1.000	1.1	1.2	inf
doggo	75.0	1.185333	0.143030	0.8	1.100	1.2	1.3	1.400000
doggofloofer	1.0	1.100000	NaN	1.1	1.100	1.1	1.1	1.100000
doggopupper	10.0	1.110000	0.228279	0.5	1.125	1.2	1.2	1.300000
doggopuppo	1.0	1.300000	NaN	1.3	1.300	1.3	1.3	1.300000
floofer	9.0	1.188889	0.105409	1.0	1.100	1.2	1.3	1.300000
pupper	224.0	1.080804	0.202534	0.3	1.000	1.1	1.2	2.700000
puppo	24.0	1.204167	0.126763	0.9	1.175	1.2	1.3	1.400000

The biggest count of dogs belong to puppers (224). The highest average rating belong to floofers (1.188889) and the lowest average rating (1.08) belong to puppers. But we can't make the conclusion, since the count is significantly different, it also states that one of the pupper has the highest maximum rating (2.7).

Correlation of favorites and count of retweets

The dataset contained information on have many retweets and favorites specific tweet received. Are you curious to see if those numbers correlates? The statistical analysis shows a large positive skewed distribution in both categories indicated by the large standard deviation. The results also proved that tweet is likely receive favorite than retweet a tweet which is shown in the plot by the larges favorite count.



According to the scatterplot above, there is a strong correlation between count of retweets and favorite count. For the visualisation I used a seaborn scatterplot.

Count of yearly retweets

Has number of yearly retweets increases from 2015 to 2017? The data in the dataset was given from 2015 to 2017, I grouped all my dates to years.

```
year
2015.0     670
2016.0    1016
2017.0     340
Name: retweet_count, dtype: int64
```

Number of retweets was increased from 670 to 1016 in 2016 and suddenly decreased to 340 in 2017. This could indicate lower popularity of the account.

Yearly average rating numerator

I used the year column that I created for the previous insight to find out if the average rating numerator has been increasing every year.

```
year
2015.0    10.235465
2016.0    13.325368
2017.0    18.052632
Name: rating_numerator, dtype: float64
```

According to my analysis based on years 2015 - 2017, it has been increasing every year. It started from 10.23 average number and increased to 18.05 in 2017.