Analyzing and Visualizing WeRateDogs Project

Presented By: Paul Anugwom

Introduction:

This report is based on findings from the analysis and visualization of the dataset from WeRateDogs. The idea is to highlight insights from the analysis and visualization.

Table1: Statistical Table

	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	followers_count	img_num	p1_conf	p2_conf	p3_conf
count	2.095000e+03	2095.000000	2095.0	2095.000000	2095.000000	2.095000e+03	1972.000000	1972.000000	1.972000e+03	1.972000e+03
mean	7.364655e+17	10.610979	10.0	2827.127446	8957.458711	3.200946e+06	1.203347	0.594026	1.348487e-01	6.018037e-02
std	6.725410e+16	2.174782	0.0	4702.022821	12198.733510	4.419017e+01	0.561517	0.272039	1.008674e-01	5.080513e-02
min	6.660209e+17	0.000000	10.0	16.000000	81.000000	3.200799e+06	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	6.765984e+17	10.000000	10.0	637.000000	2039.500000	3.200901e+06	1.000000	0.362063	5.411538e-02	1.605498e-02
50%	7.092251e+17	11.000000	10.0	1390.000000	4180.000000	3.200947e+06	1.000000	0.587797	1.184015e-01	4.947920e-02
75%	7.879251e+17	12.000000	10.0	3265.500000	11402.500000	3.201002e+06	1.000000	0.845599	1.956673e-01	9.162278e-02
max	8.924206e+17	14.000000	10.0	79515.000000	132810.000000	3.201018e+06	4.000000	1.000000	4.880140e-01	2.710420e-01

Table 2: Correlation Table

	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	followers_count	img_num	p1_conf	p2_conf	p3_conf
tweet_id	1.000000	0.518135	NaN	0.400802	0.652191	-0.865557	0.212964	0.104923	-0.002778	-0.048547
rating_numerator	0.518135	1.000000	NaN	0.307143	0.402701	-0.489331	0.192951	0.097335	0.006047	-0.024944
rating_denominator	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
retweet_count	0.400802	0.307143	NaN	1.000000	0.911221	-0.360548	0.105810	0.056346	-0.019222	-0.046288
favorite_count	0.652191	0.402701	NaN	0.911221	1.000000	-0.544332	0.135726	0.080566	-0.022856	-0.054937
followers_count	-0.865557	-0.489331	NaN	-0.360548	-0.544332	1.000000	-0.218421	-0.079836	-0.006848	0.033021
img_num	0.212964	0.192951	NaN	0.105810	0.135726	-0.218421	1.000000	0.205192	-0.157703	-0.140941
p1_conf	0.104923	0.097335	NaN	0.056346	0.080566	-0.079836	0.205192	1.000000	-0.510248	-0.709688
p2_conf	-0.002778	0.006047	NaN	-0.019222	-0.022856	-0.006848	-0.157703	-0.510248	1.000000	0.479035
p3_conf	-0.048547	-0.024944	NaN	-0.046288	-0.054937	0.033021	-0.140941	-0.709688	0.479035	1.000000

Chart 1: Correlation between Retweet Count and Favorite Count

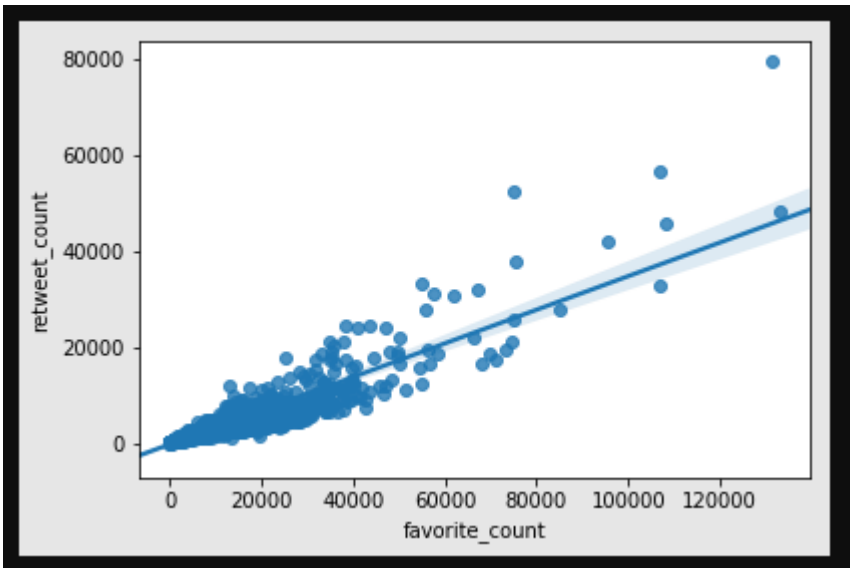


Chart 2: Histogram of Dog Names, Showing most popular dog names

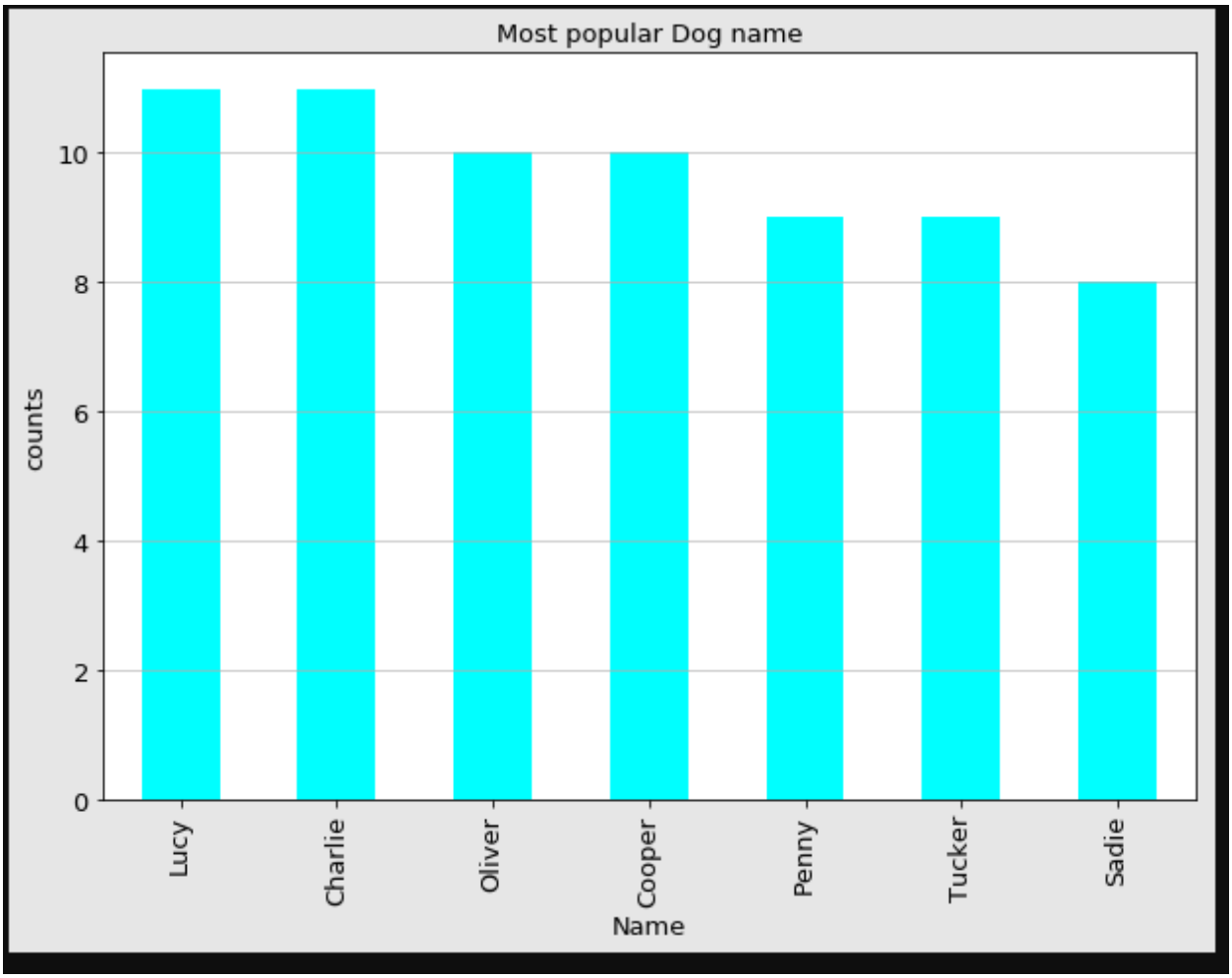
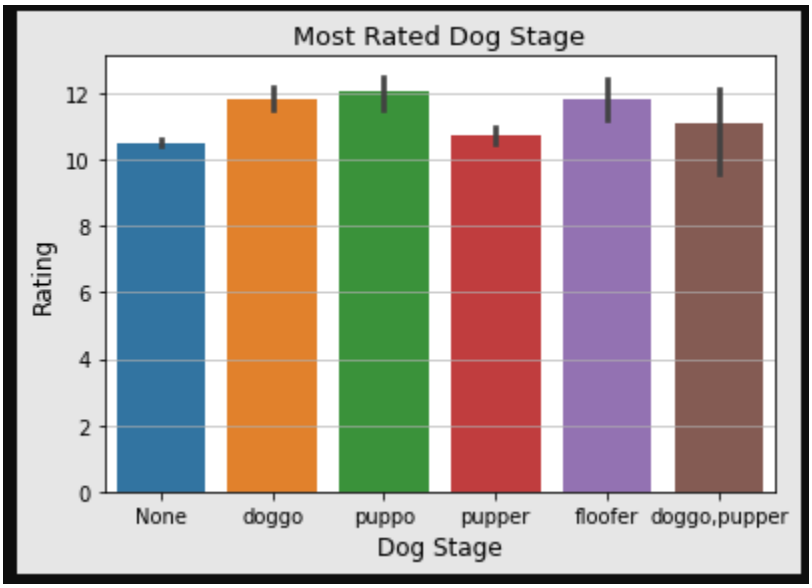


Chart 3: Bar chart showing dog stages and their Ratings



Insights from the Analysis and Visualization

1. From the Basic statistics above (table1), P1 has the highest average confidence of prediction of about 60%, which implies there is more confidence of correct prediction for p1 and least confidence of correct prediction for p3. I can deduce based on this finding that the Neural network model seems to be working fine.

2. From Chart 2 above showing the most popular dog name, I can deduce that most popular dog names are Lucy and Charlie.Also, from chart 3 above, the most rated dog stage from the visualization is Puppo.

3, From the correlation table 2, I did observe that there is a strong positive correlation between retweet count and favorite count, with a coefficient of correlation of about 0.9.This correlation between retweet count and favorite count is shown in clearly in a scatter plot(chart1) above. Another notable correlation (from table2) is between p1_conf and p3_conf.There is a moderate negative correlation between both, which implies that the more the confidence of prediction 1, the less the confidence of prediction 3.

4, Another insight worth noting is that Maximum rated dogs received a rating of 14, and 17 dogs received this rating. Most followed dogs have 3201018 followers. Gabe is the most rated dog that has most followers