Act Report

For the analysis of the final dataset, the *timestamp* column was set as the index in order to gather time series analysis. Twitter data is very time specific and the trends are nice to see visually overtime which include ‘retweet count’ and ‘favorite count.’

The first observation was done using the **describe()** method, which showed basic statistics for the numerical data in the table. Each variable initially seemed to follow a logical pattern without abnormal outliers sending red flags. For example, the **rate_denom** column, the minimum is 2 which makes sense because 0 cannot be the lowest since it's part of a ratio and you cannot divide by zero.

Even, though some of the numerators in the **rate_num** column are extraordinarily high and above 10 for the denominator, the rating system for the WeRateDogs feed follows an out of the norm rating system. Therefore, this should not draw attention as an issue.

Also, for columns **p1_conf**, **p2_conf**, and **p3_conf**, the numbers are in the bounds from 0 to 1, which is good because they are confidence intervals and are bound between 0 and 1. **p1_conf** has a higher average than **p2_conf** and so forth, and their max and min are higher in both categories respectively.

This suggests that the neural network developed to identify the breed of dog works effectively because "p1", "p2" and "p3" are the predictions in succession. Indicating the first prediction (p1) has a higher confidence of being correct than the following predictions.

The next observation was done using the **corr()** method, which finds the correlation coefficients between each variable. The correlation chart (shown below) was useful for finding connections between variables, especially with hypothesis testing or an A/B test. The numbers range from 0 to 1, and a positive number is a positive correlation and vice versa for a negative number.

The correlation coefficient between **retweet_count** and **favorite_count** is 0.929717, which is close to 1 and positive demonstrating a strong positive correlation between those two metrics.

The only other coefficient worth mentioning is the between **p1_conf** and **p3_conf** which is -0.707485, demonstrating a negative semi-strong relationship. This is probably a result because **p3_conf** confidence score is affected by the results of the first

prediction in the neural network. The more confident the result of the first image, the less likely the following images would accurately guess the breed of dog.

The last observation was done by plotting time series trend lines of the *retweet_count* and the *favorite_count*. The line graph (shown below) was created by resampling the average counts of data, in weekly intervals, much like a moving average, which helps smooth out the graph and improve visibility of the trends.

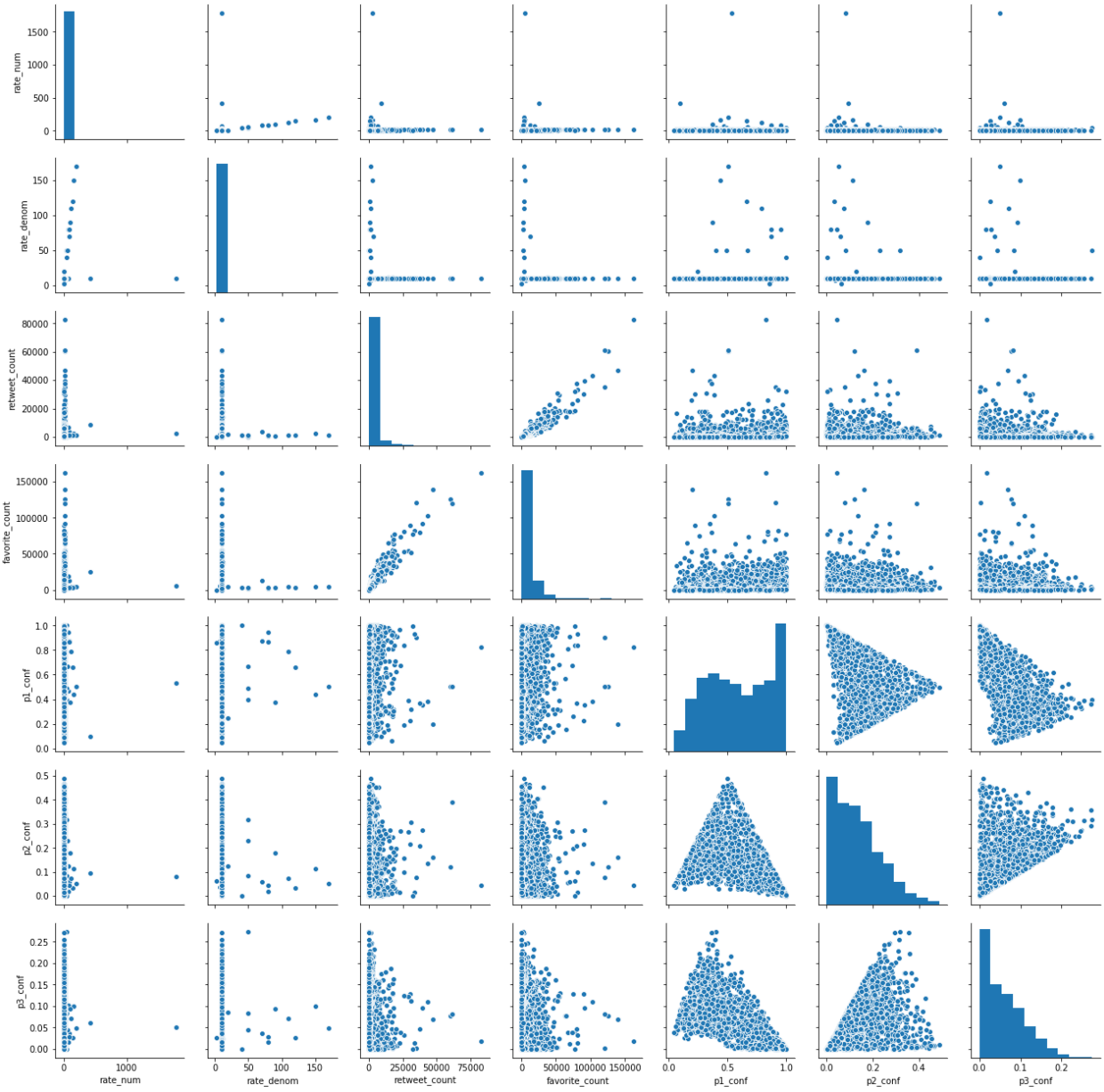
As mentioned before, with this timeseries chart, the **favorite_count** and **retweet_count** are positively correlated with one another. This is due to the fact that most people retweet 'tweets' that they like in order for others to see it. It is like 'free advertising' for the 'tweet' itself and shows people on your feed what you're interested in. It's a sharing feature when you want others to see something you've read/seen.

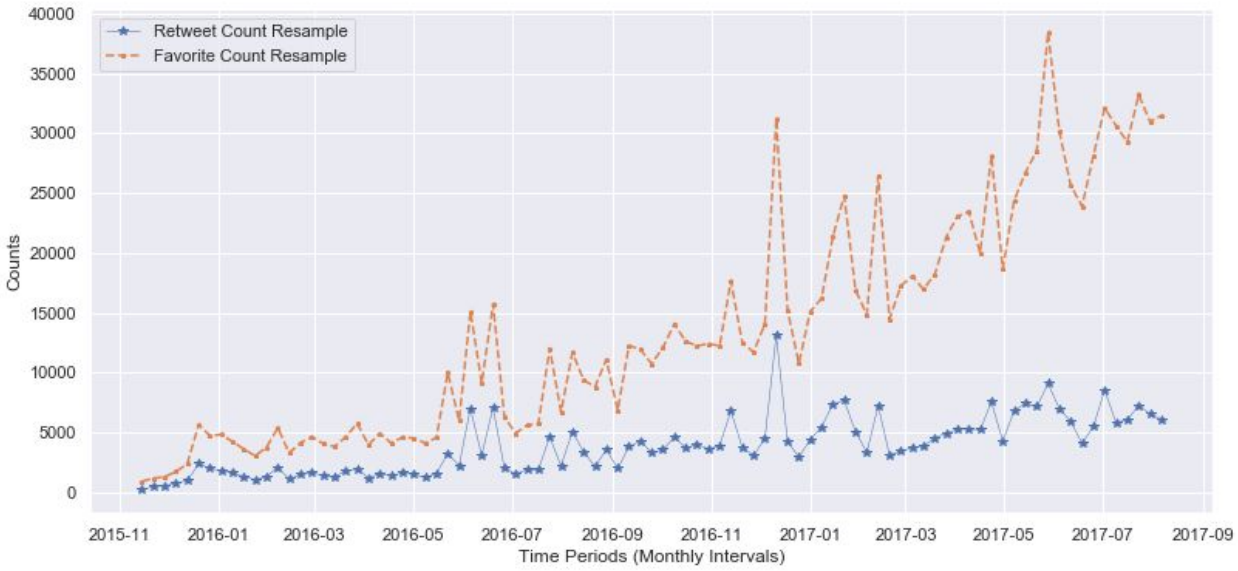
According to the chart, there are 3 spikes at specific times of year, those being, the middle of spring into summer, and then the Christmas/Holiday time. This is most likely due to the fact that people with dogs are more active in the warmer months, posting cute things their dog is doing outside. Also, during the holidays, they are more likely to share pics/etc about things they care about during this time i.e. their dogs.

The overall trend of the counts shows the popularity of this twitter page slowly growing overtime. If the number of followers were taken into account, it would most likely show a similar trend since the twitter account is getting more "air time."

Finally a fun way to see the most used words in the tweets, was to create a word cloud (shown below). A word cloud is a fun tool that lets user take the most frequently used words from a text, in this case, the tweets used in the dataset, and display in a fun image. The outline of a paw print was used for this word cloud in association with the @WeRateDogs twitter account that was used in the process of this project.


Meet
pupper
af
dog
pup
doggo
Love
great
want
know
tho
amp
will
look
ckin
good
take
now
pet
really
still
time
got
Say
hello
happy
one
day
quite
rate dog
still
trying
rare
well
ck
eye
first
give
tongue
Thank
probably
enjoy
need
much
see
think
back
someone
even
go
help
hug
nifty
good boy
guy
hell
owner
tongue slip
fluffy
damn
never
pretty
going
many
hat
super
Always
send dogs
likes
long
Please send
dogs Please
little
stop sending
Christmas
thing
keep
snug
two
puppo
today
Cool
ear
appear
work
make
Seem
selfie
better
pic
adorable
favorite
tragic
extremely
looking
lil
nice
Charlie
cuddle
flower
Must
friend
proud
clearly


rate_num
rate_denom
retweet_count
favorite_count
p1_conf
p2_conf
p3_conf


Retweet Count Resample
Favorite Count Resample
Counts
40000
35000
30000
25000
20000
15000
10000
5000
0
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
Time Periods (Monthly Intervals)