Report by Anas A. Mohaisin

Post-Data Wrangling

As a starting point we visualise the distribution & the probability density function (PDF) of Retweet and Favourite. Figure 1 shows both variables are positively skewed which means that the retweet and favourite counts are less than the average 2976.08 and 8556.71, respectively. Kurtosis is platykurtic which means the peak is lower and broader than Mesokurtic

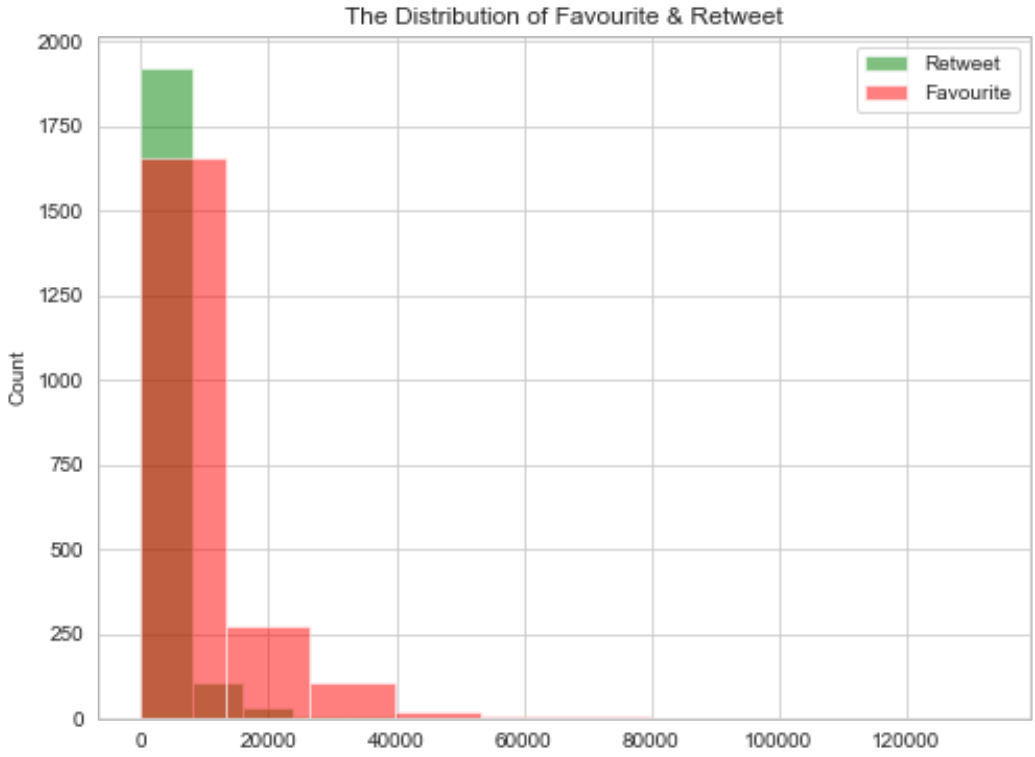

Figure 1: shows the distribution of Favourite and Retweet from the WeRateDogs cleaned dataset.

Figure 2, shows a platykurtic *cf.* mesokurtic where the . PDF is the function of the probability; i.e.

$$f(X) \geq 0,\ so\ the\ probability\ for\ a\ and\ b\ P(a \leq X \leq= \int_a^b f(X)dx.$$

In the other words, this figure shows the higher and the lower probability of a drawn sample from the Retweet and Favourite data.

Theoretically, the more you tweet, the more you like; Figure 3 shows highly positive correlation between the Favourite and Retweet 79%. This figure also shows how the data are dispersed which does not show high dispersion. However, one of the statistical methods to measure the dispersion is the variance. We shall plot a linear regression curve as figure 4.

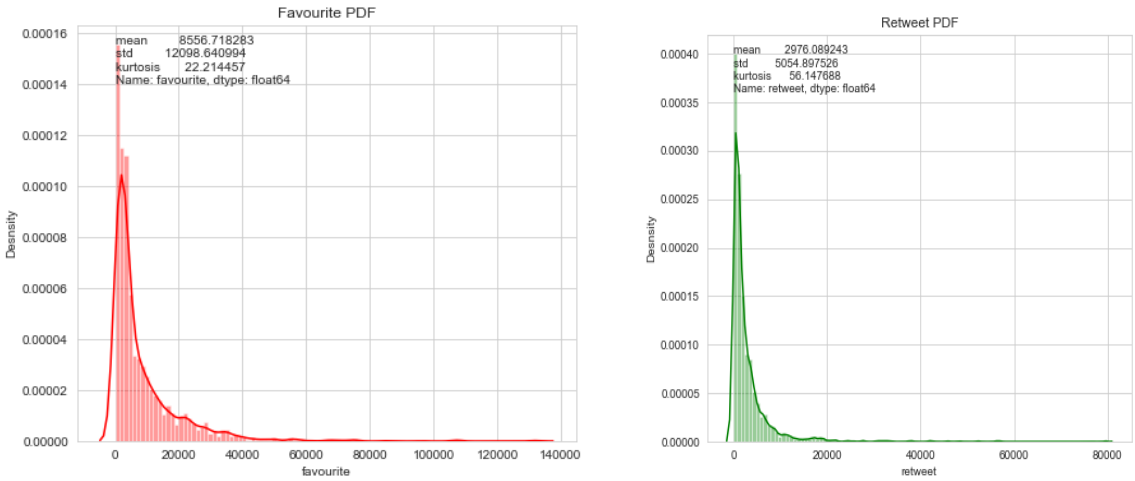

Figure 2: Shows the PDF for both variables from the WeRateDogs cleaned dataset

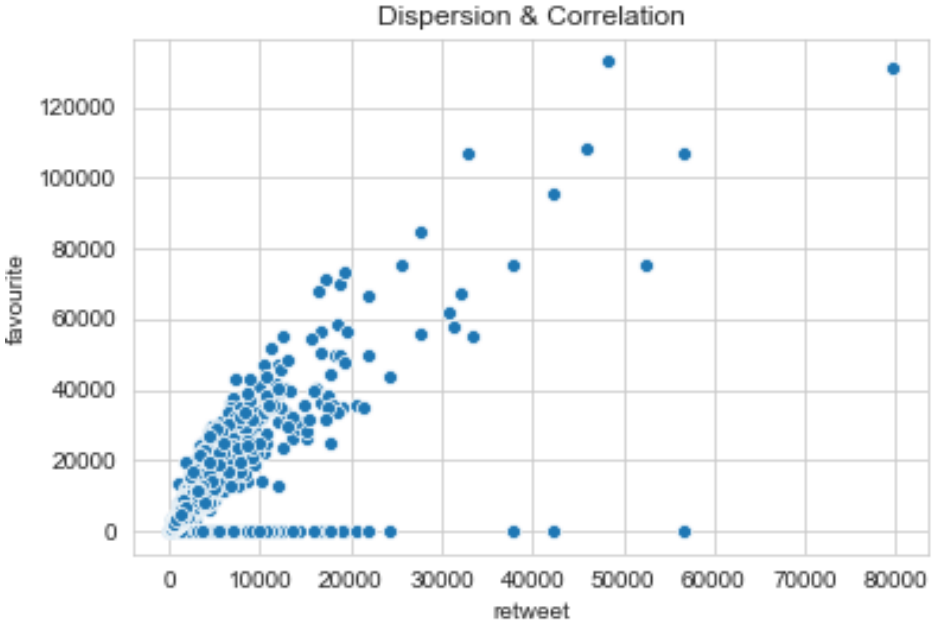

Figure 3: Scatter plot shows the dispersion and correlation for Favourite *vs* Retweet

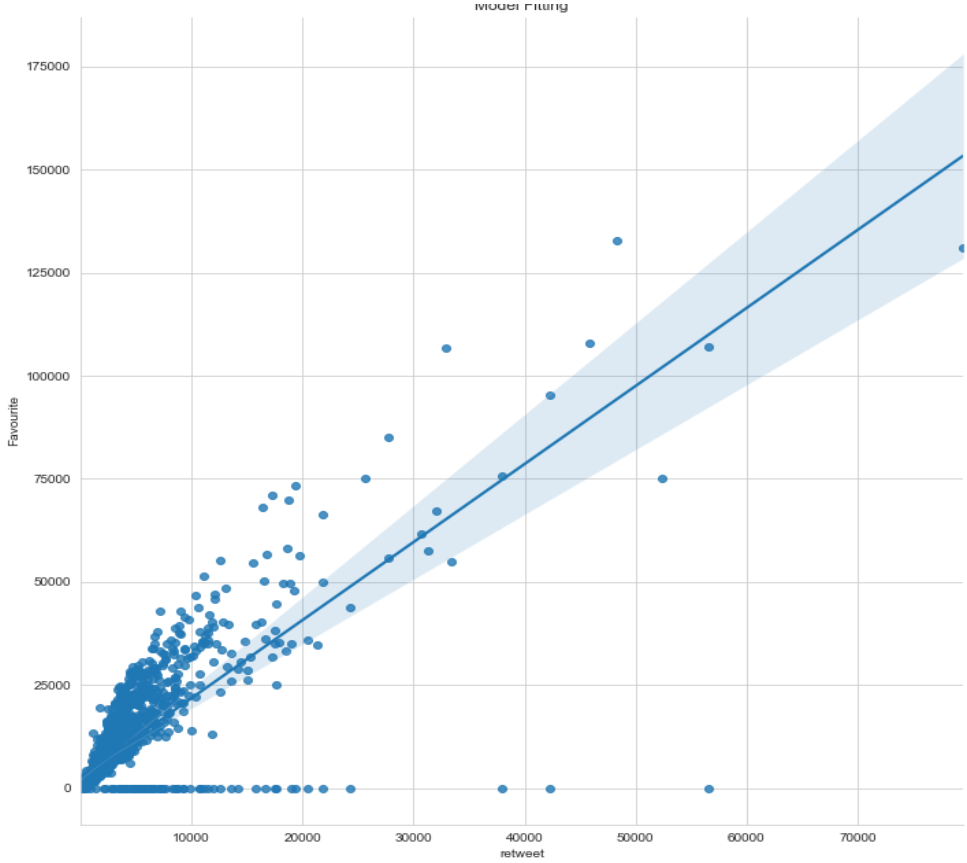

Figure 4: Fitting the favourite and retweet model

We assume that the Favourite is a function in Retweet. Thus, the simple regression model[1] found that an increase in Retweet by one unit, increase the Favourite by 1.89 which is statistically significant at $\alpha = 0.05$ et ceteris paribus. One of the OLS model assumptions is a constant variance; i.e. *Homoscedasticity*. Breusch-Pagan test for the homoscedasticity found that we accept the $H_0: .Var(\epsilon) = E(\epsilon^2|x_i) = \sigma^2$ which means the variance of the conditional error term is homoscedastic. [2]

Dog stages data can be visualised by using a pie chart with relevant percentages as per figure 5. It shows that Pupper's dog stage has the highest proportion 69.1% among other stages while Doggo comes next in 20.9%. The rest of proportions are unequally distributed between Puppo and Flooter.[3]

Crosstab analysis for Dog stages for counts and means shows the dog states -including None- has same count of Retweet and Favourite with different means.

Figure 6 shows the proportions of source type where Twitter from the iphone source has the highest proportion 98%.

[1] Just for the sake of explanation
[2] Normality assumption is relaxed as per Lagrange Multiplier (IID) and F-test
[3] We have dropped the None from dog stages pie chart.

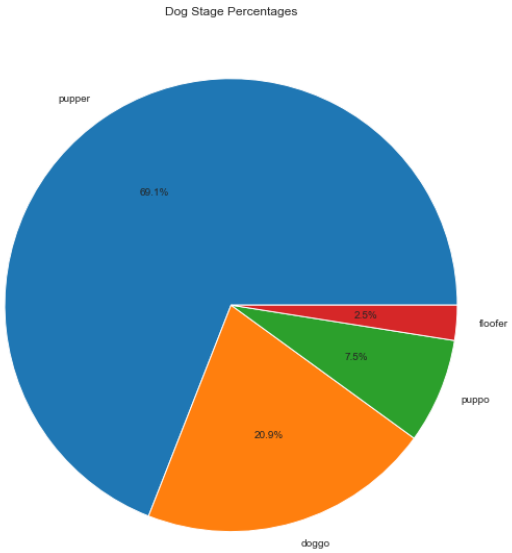

Figure 5: Pie plot shows the dog stages proportions.

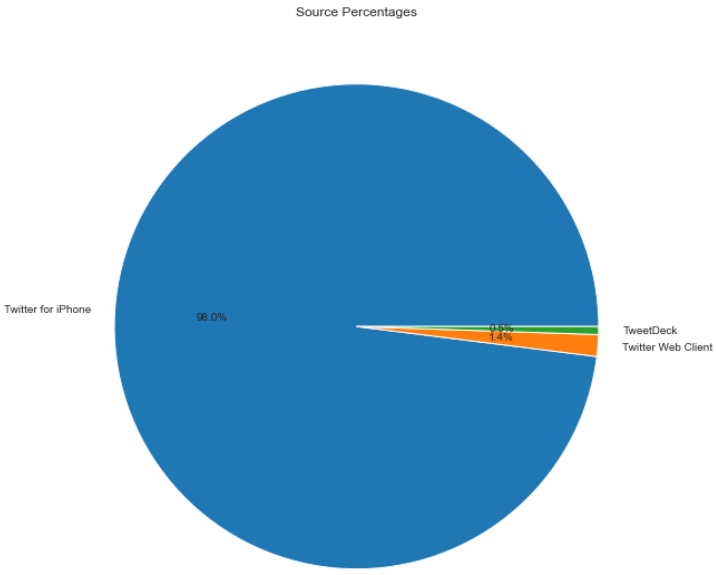

Figure 6: Source type proportions

Figure 5 shows how the dog stages are dispersed among Retweet and Favourite data and how are they fit. It shows different intercept points based on the stages of dogs, but the dispersion is the same as Figure 4.

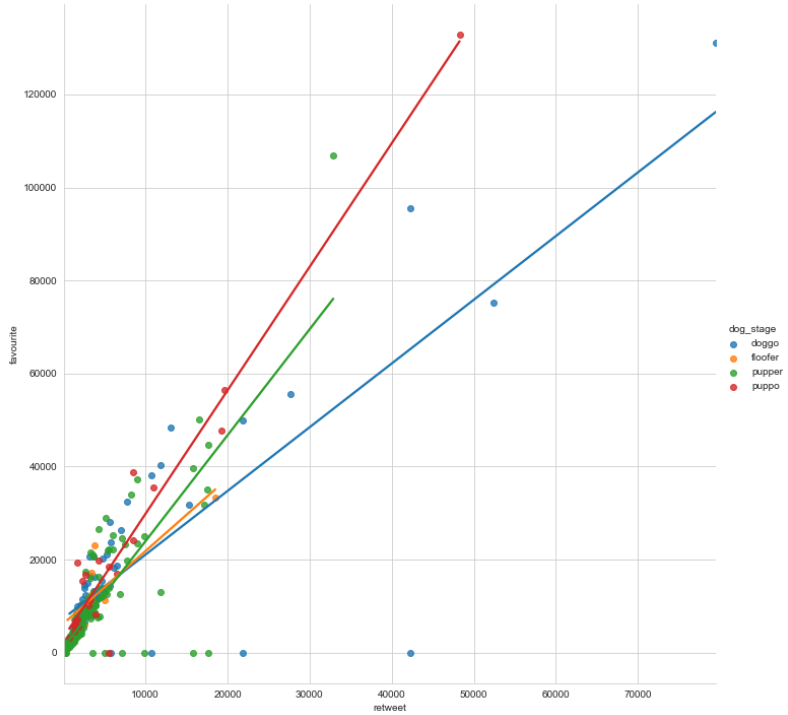

Figure 7: Fitting model for the dog stages, favourite, and retweet data

The next part of our analysis tried to shed the light on the dimension of time[4] with respect of Retweet and Favourite variables. By comparing the monthly tweet and favourite data, we found that December has the highest total 897,905 and 216,6054, respectively. Despite of having two Decembers in our dataset the total number exceeded other months; e.g. Christmas time and the New Year. People who buy like pets may prefer to buy/sell the highly favourite and tweeted during the Christmas time.

The above analysis leads us to explore the data over time and to test whether the time is a factor of the Retweet or/and Favourite or they are correlated with respect to time. Based on our data 2105/11-2017/08, we cannot investigate the seasonality because we do not have at least eight quarters or twenty-four months. So, we shall investigate the total monthly Retweet and Favourite processes as Figure 8. As we see- both processes have similar troughs and peaks over time, which is *not*, necessarily, mean that they are correlated e.g. spurious[5].

[4] We shall proceed the very basic time-series analysis
[5] We shall not process with it because we do not have adequate variables to and time-series.

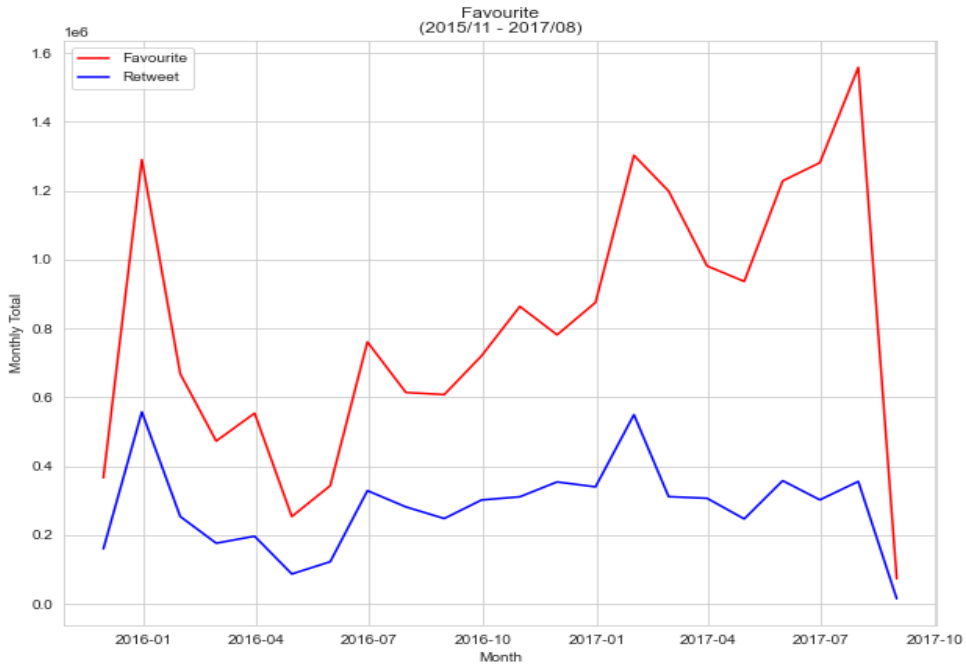

Figure 8: Total Retweet and Favourite time-series on monthly basis.

Visualisation of the Log transformation is an effective method to stabilise the variance over time as in Figure 9.

As a starting point we need to test the processes for stationarity. Mathematically, we the process should be:

$$E(x_t) = \mu$$
$$Var(x_t) = \sigma^2$$
$$Cov(x_t, x_{t-1}) \neq f(t)$$

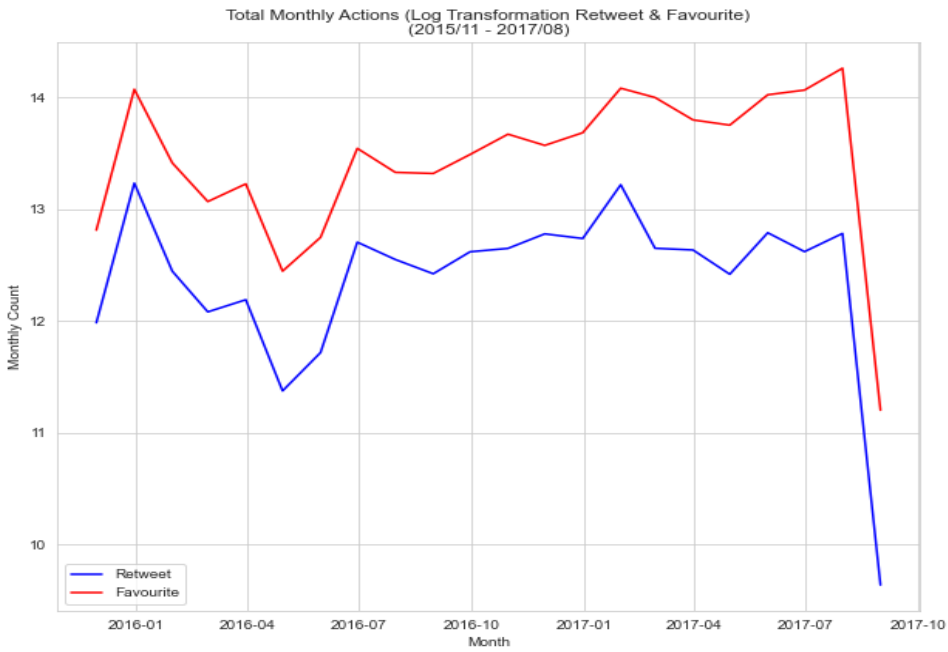

Figure 9: Log-transformation for total Retweet and Favourite time-series on monthly basis

Augmented Dickey–Fuller (ADF) test for the stationarity:

- We fail to accept null hypothesis that the Retweet process has a unit root; i.e. $\rho = 1,\ where\ \Delta X_t = \mu + (\rho - 1)X_{t-1} + e_t$.[6]
- We accept null hypothesis that the Favourite process has a unit root; i.e. $\rho = 1,\ where\ \Delta X_t = \mu + (\rho - 1)X_{t-1} + e_t$.[7]
- We conclude that the Favourite has a stationary process while the Retweet is non-stationary.

[6] ADF Stat: -3.666262, p-value: 0.004614
[7] ADF Stat: -1.206518, p-value: 0.67078