act_report

May 22, 2022

0.1 Report: act_report

0.1.1 Insights and Visualisation

The We Rate Dogs twitter handle was wrangled, analysed and visualised. Some insights has been gained on how some certain variables have impact other variables and it affects tweeters from this account.

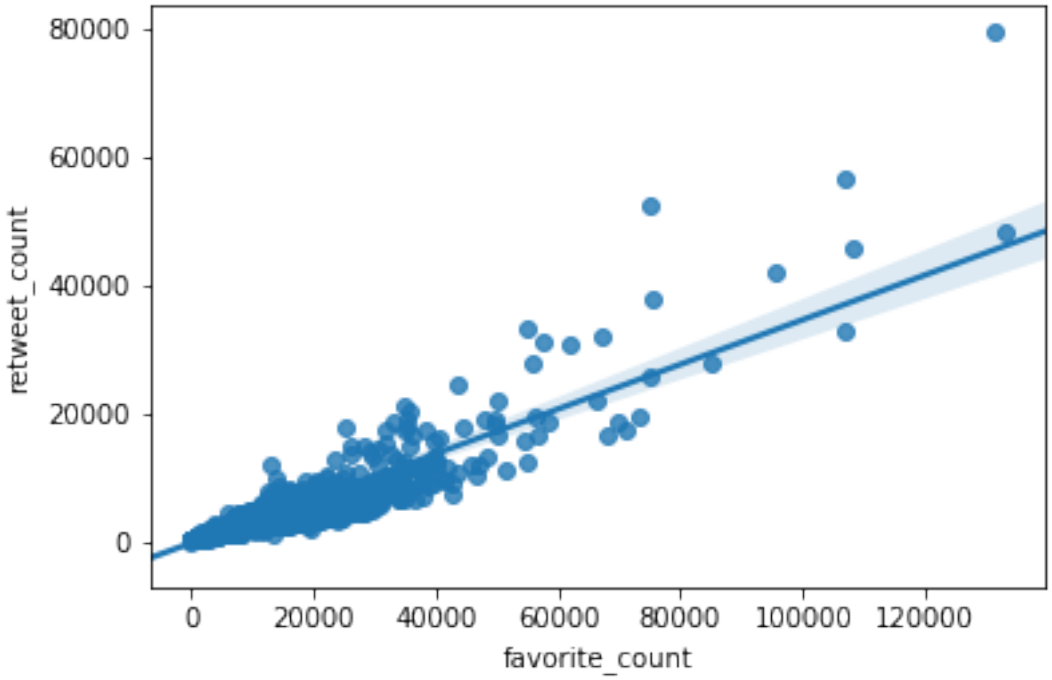
```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
```

```
In [2]: master = pd.read_csv('twitter_archive_master.csv')
```

1. There is seems to be a strong positive correlation between retweet_count and favorite_count. Both variables move in the same direction, that means the change in one variable will affect the change in the other. The higher the favorite_count the higher the retweet_count and vice versa.

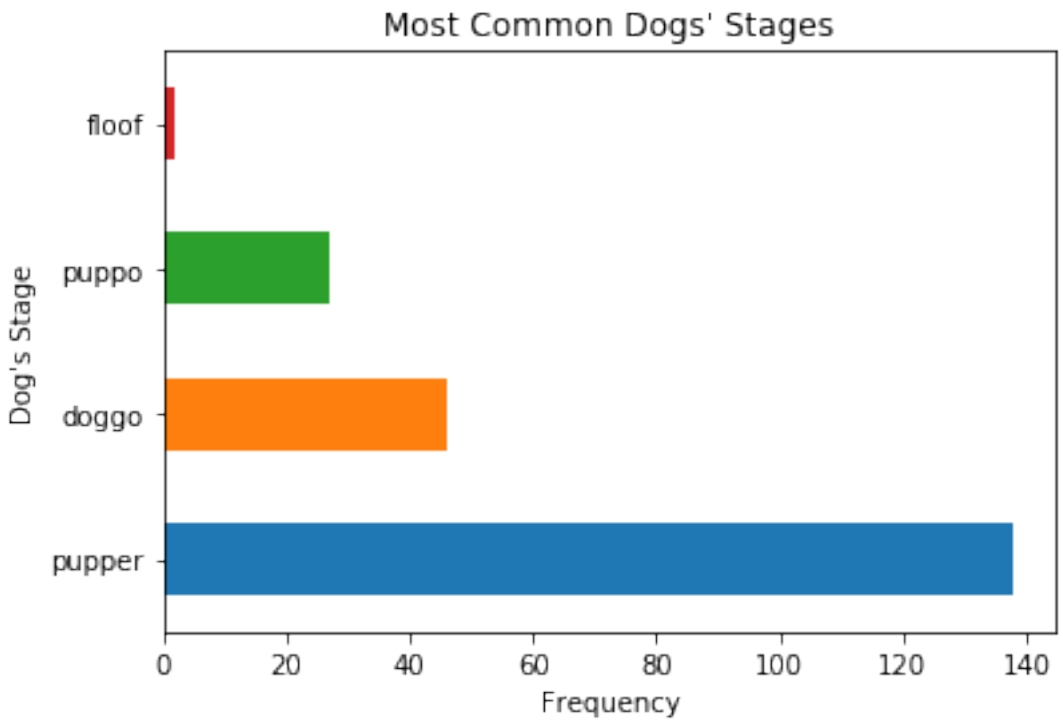
```
In [3]: sns.regplot(x=master['favorite_count'], y=master['retweet_count'])
```

```
Out[3]: <matplotlib.axes._subplots.AxesSubplot at 0x7f49d5125b70>
```



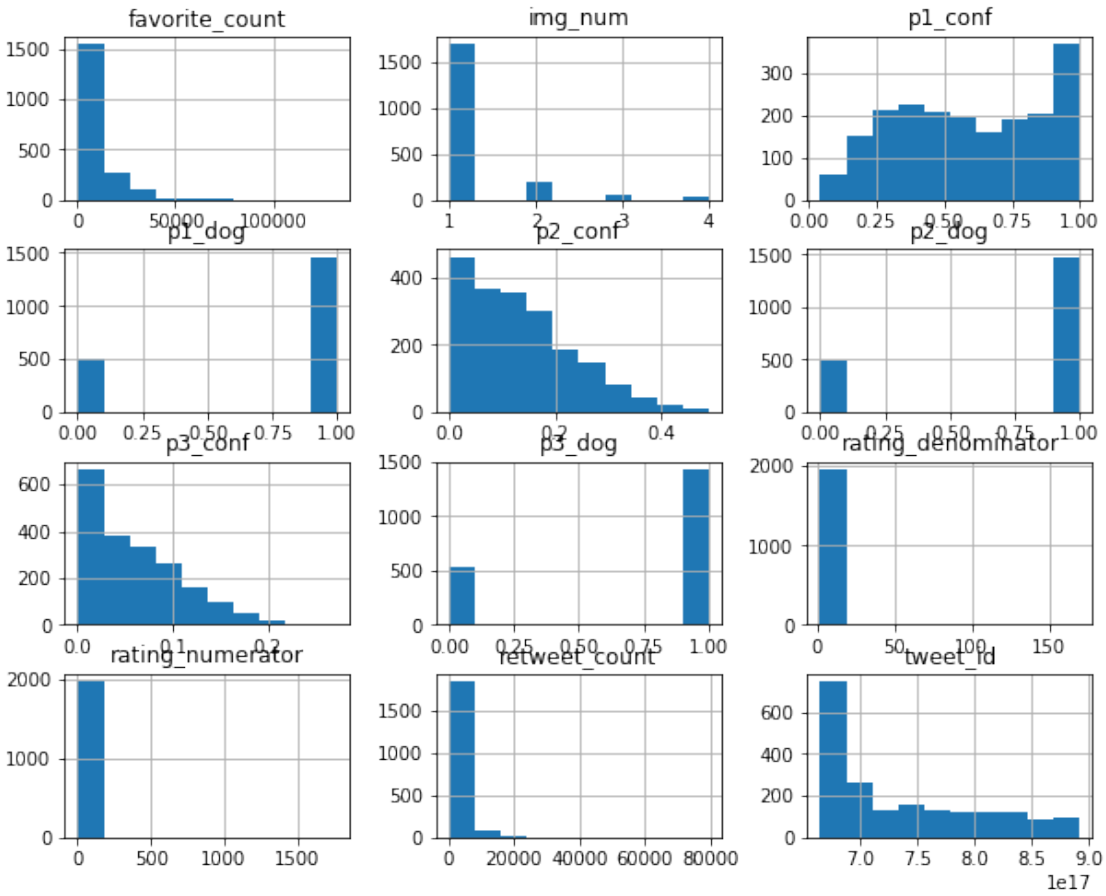
2. The Dog stage with the highest frequency is the Pupper stage, followed by the other dog stage doggo, puppo and lastly floof. Pupper is the most dog stage occurring more than half of the other dog stages.

```
In [4]: # rank the stages frequency in a descending order
        master.dog_stage.value_counts().sort_values(ascending =False)[:10].plot(kind = 'barh')
        plt.title("Most Common Dogs' Stages")
        plt.xlabel('Frequency')
        plt.ylabel("Dog's Stage");
```



3. The numerical variables are right-skewed distribution, and that means there is a positive relationship that is, the mean is greater than the median. The others are left-skewed distribution, and that means there is a negative relationship, that is, the median is greater than the mean.

```
In [5]: master.hist(figsize=(10,8));
```



4. The most common name is a bit inaccurate because most of it were none occurring about 524 and that is a lot. The next was the letter a as the second most common name, it was probably mistaken as the dog's name instead of an article.

```
In [6]: from collections import Counter
        x = master['name']

        count = Counter(x)
        count.most_common(11)
```

```
Out[6]: [('None', 524),
         ('A', 55),
         ('Charlie', 11),
         ('Oliver', 10),
         ('Cooper', 10),
         ('Lucy', 10),
         ('Penny', 9),
         ('Tucker', 9),
         ('Sadie', 8),
```

```
('Winston', 8),
('Lola', 7)]
```

5. The most commonly used tweets' source from the total tweets is Twitter for iPhone, making it look like the other sources (Twitter web client and tweetDeck) are non-existent because of the amount of tweeters use that Twitter for iPhone. This might be as a result of the demographic or the region of the tweeters.

```
In [7]: plt.title("Distribution of Tweets'Source")
        master.source.value_counts().sort_values().plot(kind ='barh')
        plt.xlabel('Total Tweets')
        plt.ylabel('Source');
```