In WeRateDogs project wrangling, we took a lot of time to gather, assess and clean the data sets. After putting in a lot of effort to get a clean dataset, we went further to generate some insights regarding the data. To produce some insights, we carefully went trhought the clean dataset to see what information we can derive out of it. The following insights were drafted out and they are listed thus:

1. Learn which rating has the highest retweet count and favorite count

2. Learn which dog has the highest rating

3. Learn what month has most retweet count and favorite count

4. Which prediction has the best confidence level

5. Learn which day has most retweeet count and favorite count

To visualize the generated insights we did the following:

We created two functions to plot our graph

\*\*Function 1\*\*: This functions plots a graph with only the x-axis column name give as an argument for x\_val parameter

def plot\_graph\_with\_xval(df, x\_val, title\_val):

df.plot(kind='bar', x=x\_val, title=title\_val, figsize=(12,12))

plt.xlabel(x\_val)

#plt.ylabel(y\_val)

plt.show(block=True);

\*\*Function\*\*: This functions plots a graph with both the x-axis and y-axis column name give as an argument for x\_val and y\_val parameter respectively

def plot\_graph\_with\_xval\_yval(df, x\_val, y\_val, title\_val):

df.plot(kind='bar', x=x\_val, y=y\_val, title=title\_val, figsize=(12,12))

plt.ylabel(y\_val)

plt.xlabel(x\_val)

plt.show(block=True);

**Insight 1:**

We got the rating and the corresponding sum of the retweets and favorite count of each associated dog. The code is given below;

We\_rate\_dogs\_insight1 = We\_rate\_dogs\_final.groupby(["rating"],as\_index=False)["retweet\_count", "favorite\_count"].sum()

We\_rate\_dogs\_insight1.sort\_values(by=["retweet\_count"], ascending = False).head(7)

The above displayed a table with the rating of 1.3 have the highest retweet count and favourite counts of 1888481 and 5601289 respectively.

|  | rating | retweet\_count | favorite\_count |
| --- | --- | --- | --- |
| 13 | 1.3 | 1888481 | 5601289 |
| 12 | 1.2 | 1515038 | 4659641 |
| 11 | 1.1 | 962155 | 2499888 |
| 10 | 1.0 | 604606 | 1518405 |
| 14 | 1.4 | 365800 | 1002910 |
| 9 | 0.9 | 119422 | 352191 |
| 8 | 0.8 | 82971 | 201068 |

Table 1.0: table for rating and their repective sum of retweet and favorite count

We call the first function and our visualization was plotted

# Call the plot\_graph\_with\_xval to plot the needed graph

plot\_graph\_with\_xval(We\_rate\_dogs\_insight1,'rating', 'Count per Rating')

Our visualization also clearly showed that 1.3 rating had the highest value of retweet and favorite counts.

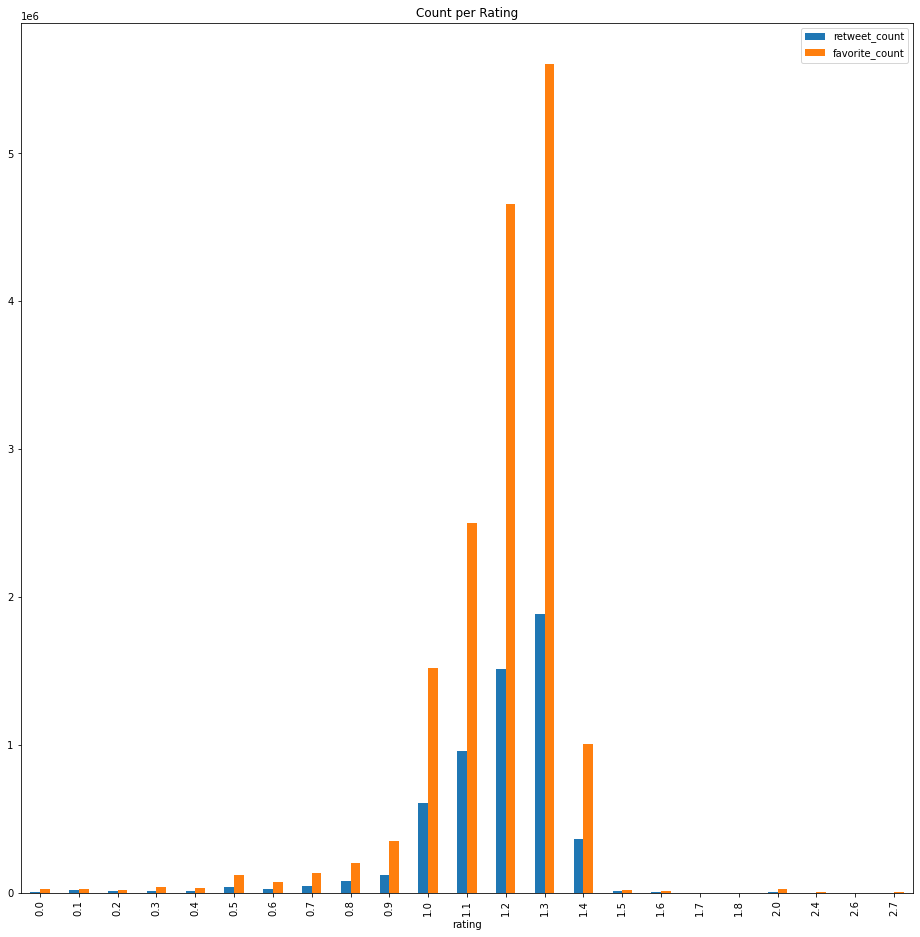


Fig 1.0: visualization for rating with the highest value

**Insight 2:**

We got the dog name and the corresponding sum of the retweets and favorite count of each associated dog thus;

We\_rate\_dogs\_insight2 = We\_rate\_dogs\_final.groupby(["name"],as\_index=False)["retweet\_count", "favorite\_count"].sum()

We\_rate\_dogs\_insight2.sort\_values(by=["retweet\_count"], ascending = False).head(7)

The above code snippest displayed the result below;

|  | name | retweet\_count | favorite\_count |
| --- | --- | --- | --- |
| 625 | none | 2089517 | 5300027 |
| 136 | buddy | 60668 | 65704 |
| 841 | sunny | 55120 | 76711 |
| 824 | stephan | 51663 | 111681 |
| 777 | seamus | 38904 | 40204 |
| 261 | duddles | 37403 | 92778 |
| 392 | hurley | 34056 | 29431 |

Table 1.1: table for dog name and their repective sum of retweet and favorite count

‘None’ stands for dogs without name, as such we won’t say it is a dog. Rather we will point out that ‘buddy’ is the dog with the highest retweet and favorite counts of 60668 and stephan' has the highest favorite count of 111681

**Insight 3:**

We got the tweet\_month and the corresponding sum of the retweets and favorite count of each associated dog.

We\_rate\_dogs\_insight3 = We\_rate\_dogs\_final.groupby(["tweet\_month"],as\_index=False)["retweet\_count", "favorite\_count"].sum()

We\_rate\_dogs\_insight3.sort\_values(by=["retweet\_count"], ascending = False).head(5)

This result is displayed thus;

|  | tweet\_month | retweet\_count | favorite\_count |
| --- | --- | --- | --- |
| 2 | December | 795308 | 1999699 |
| 4 | January | 758681 | 1814609 |
| 6 | June | 608686 | 1990703 |
| 5 | July | 568128 | 2063657 |
| 3 | February | 485770 | 1477579 |

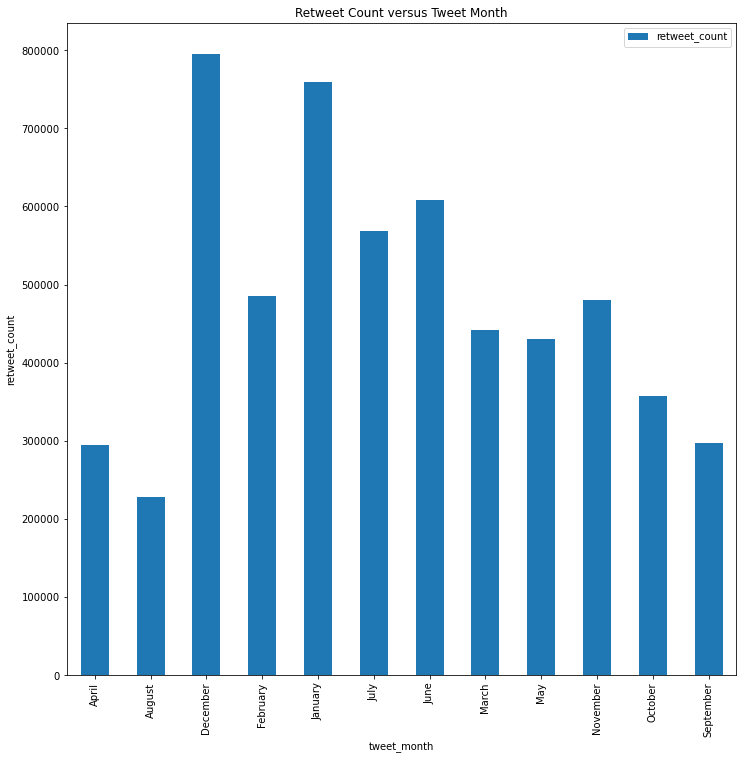
Table 1.2: table for tweet month and their repective sum of retweet and favorite count

December possesses the highest retweet while July possesses the highest favorite count.

Our visualization for this insight was plotted thus;

#call plot\_graph\_with\_xval\_yval to plot the graph

plot\_graph\_with\_xval\_yval(We\_rate\_dogs\_insight3, 'tweet\_month', 'retweet\_count', 'Retweet Count versus Tweet Month')

Fig. 1.1 Retweet count per tweet month

From the visualization above most of the retweet was done in the month of December with January as a second. I assumed that this could be because of the christmas and new year holidays.

The visualization below, clearly shows July with the highest favorite count.

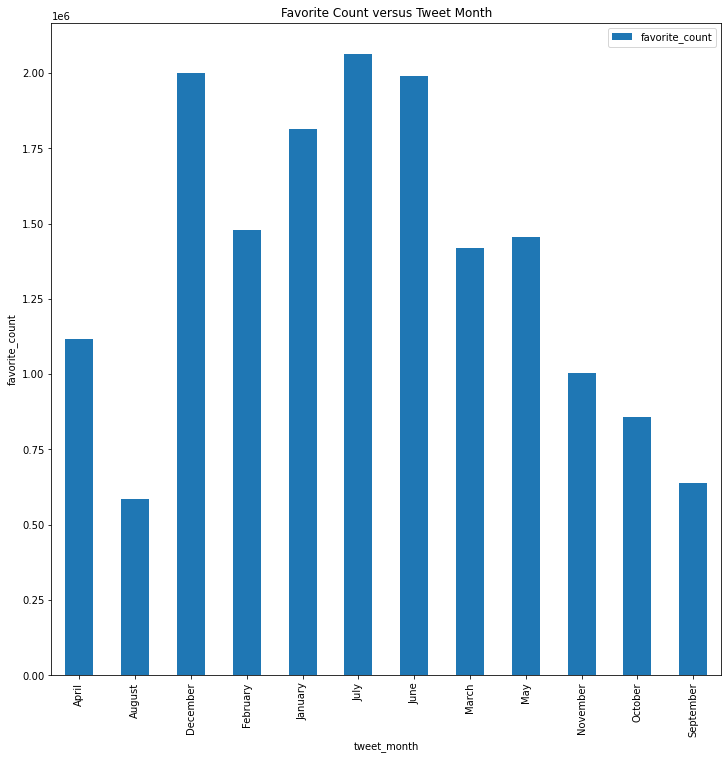


Fig. 1.2 Favorite count per tweet month

**Insight 4:**

We got the descriptive summary using;

We\_rate\_dogs\_final.describe()

|  | tweet\_id | p1\_conf | p2\_conf | p3\_conf |  |
| --- | --- | --- | --- | --- | --- |
| count | 2.327000e+03 | 1232.000000 | 1232.000000 | 1232.000000 |  |
| mean | 7.417930e+17 | 0.626173 | 0.143554 | 0.061857 |  |
| std | 6.820795e+16 | 0.253180 | 0.103855 | 0.053140 |  |
| min | 6.660209e+17 | 0.044333 | 0.000056 | 0.000008 |  |
| 25% | 6.781394e+17 | 0.414978 | 0.055919 | 0.015925 |  |
| 50% | 7.178418e+17 | 0.627879 | 0.131007 | 0.048775 |  |
| 75% | 7.986547e+17 | 0.853315 | 0.207230 | 0.095678 |  |
| max | 8.924206e+17 | 0.999885 | 0.467678 | 0.273419 |  |

Table 1.3: Descriptive summary of the dataframe

Since the highest value of a confidence level is 1. On the max row, P1\_conf has the best value (0.999885) close to 1. Also on the 50 percentile, p1\_conf has the best value

**Insight 5:**

In learning which day has the most retweet and favorite count, we got the day and the sum of the retweet and favorite count as thus;

We\_rate\_dogs\_insight5 = We\_rate\_dogs\_final.groupby(["tweet\_of\_day\_of\_week"],as\_index=False)["retweet\_count", "favorite\_count"].sum()

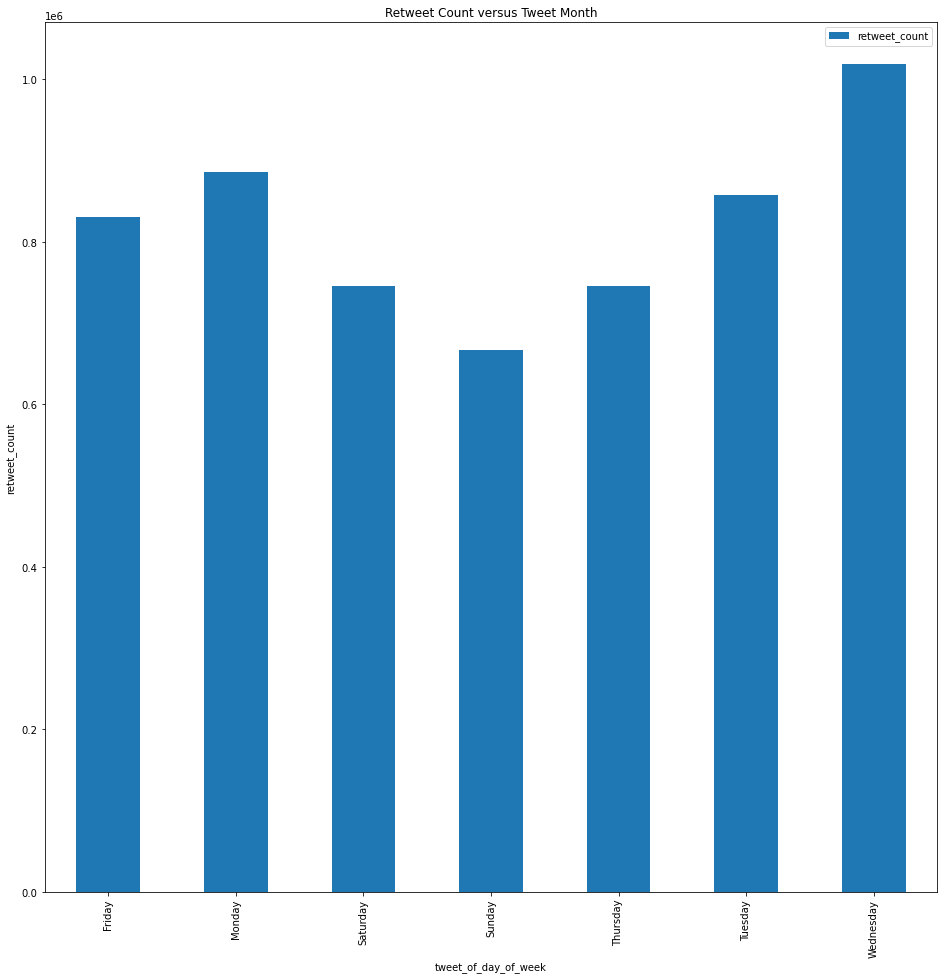
We\_rate\_dogs\_insight5.sort\_values(by=["retweet\_count"], ascending = False).head(5)

The result of the above code cell is given below;

|  | tweet\_of\_day\_of\_week | retweet\_count | favorite\_count |
| --- | --- | --- | --- |
| 6 | Wednesday | 1018794 | 2693812 |
| 1 | Monday | 885853 | 2631188 |
| 5 | Tuesday | 857425 | 2481202 |
| 0 | Friday | 830125 | 2240704 |
| 4 | Thursday | 745045 | 2102115 |

Table 1.4: table for tweet day of the week and their repective sum of retweet and favorite count

The visualization plot below shows Wednesday having the highest retweet count



The next visualization also shows wednesday having the highest favorite count

