Data Analysis And Visualization

Introduction

This report includes the data visualization after we cleaned the data. This visualization helps to prove insights about about our analysis and try to give picture of our analysis.

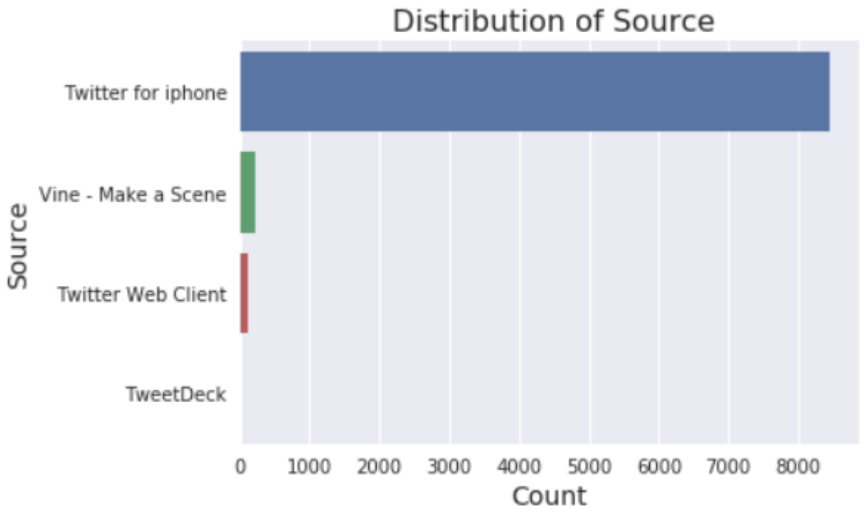
Data Analysis and Visualization

1. **Distribution of source**

One of the first visualization that I performed was to know the distribution of the source that twitter feed is getting. Looking at the chart, We can clearly say that about majority ~93% of the tweets are coming from iPhone while tweet deck source is kind of rare.

```
Twitter for iphone     8437
Vine - Make a Scene     208
Twitter Web Client      105
TweetDeck                11
Name: source, dtype: int64
```

```
[231]: Text(0.5,1,' Distribution of Source')
```



2. Histogram of dog type

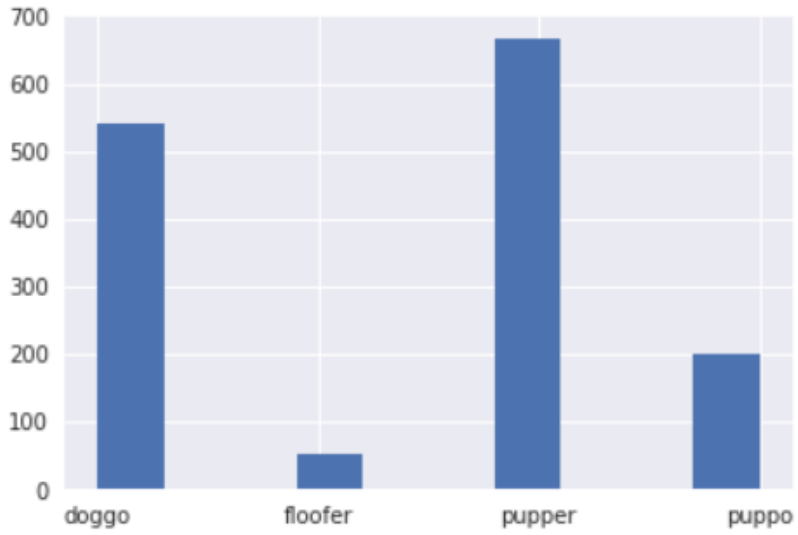
Second visualization indicates about the type of dogs that has been indicated in the tweets. Looking at histogram, we can clearly see that pupper is the highest mention in the tweet and floofer is the lowest mention in the tweet.

```
In [240]: tarchive.dogtype.value_counts()

Out[240]: pupper     667
          doggo      542
          puppo      200
          floofer     50
          Name: dogtype, dtype: int64

In [241]: tarchive.dogtype.hist()

Out[241]: <matplotlib.axes._subplots.AxesSubplot at 0x7feb2f039908>
```



3. Plot of retweet that has less count than 20000

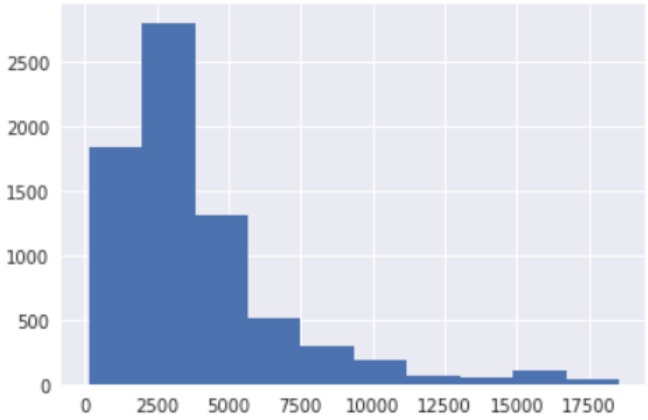
In this visualization, histogram shows about the tweet that has the retweet count less than or equal to 20000. Looking at the visualization and data we can see that the histogram is right skewed.

```
In [243]: tarchive.retweetcount.describe()

Out[243]: count     7360.000000
          mean      4486.392391
          std       4965.260908
          min        146.000000
          25%       2005.000000
          50%       3080.500000
          75%       4876.250000
          max      54120.000000
          Name: retweetcount, dtype: float64
```

```
In [224]: tarchive[tarchive.retweetcount<=20000].retweetcount.hist()

Out[224]: <matplotlib.axes._subplots.AxesSubplot at 0x7feb2fa47940>
```



4. Plot of favorite count of the tweet

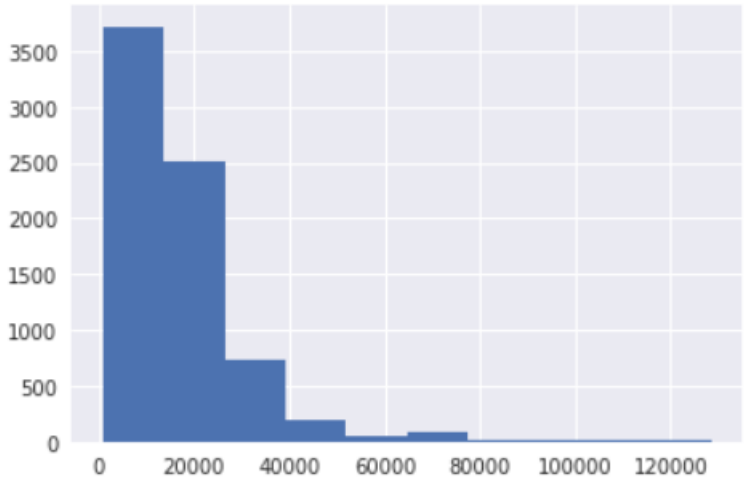
In the histogram and looking at the record information we can clearly see that it is right skewed. Also, mean is greater than median so it write skewed.

```
In [225]: tarchive.favcount.describe()

Out[225]: count      7360.000000
          mean      17326.679484
          std       14086.175013
          min         843.000000
          25%        8528.000000
          50%       13491.000000
          75%       21717.250000
          max      128674.000000
          Name: favcount, dtype: float64

In [226]: tarchive.favcount.hist()

Out[226]: <matplotlib.axes._subplots.AxesSubplot at 0x7feb304e28d0>
```



5. Highest name that has been used in twitter

Looking at the descriptive value that has the highest number of names that been used in the tweet.

```
In [227]: tarchive.name.value_counts().head(20)
```

```
Out[227]: None       2005
          a            84
          Charlie      68
          Penny        57
          Tucker       55
          Cooper       48
          Stanley      47
          Daisy        46
          Winston      45
          Bailey       44
          Bo           42
          Oliver       40
          Lucy         39
          Boomer       37
          Koda         35
          Scout        34
          Zeke         34
          Toby         34
          Leo          34
          Rusty        34
          Name: name, dtype: int64
```

6. Correlation between retweet count and favorite count

I just want to test the correlation between retweet count and favorite count. Looking at the visualization we can clearly see that it is positive correlation. It does have strong association between this variable.

```
Out[228]: <matplotlib.axes._subplots.AxesSubplot at 0x7feb30117470>
```