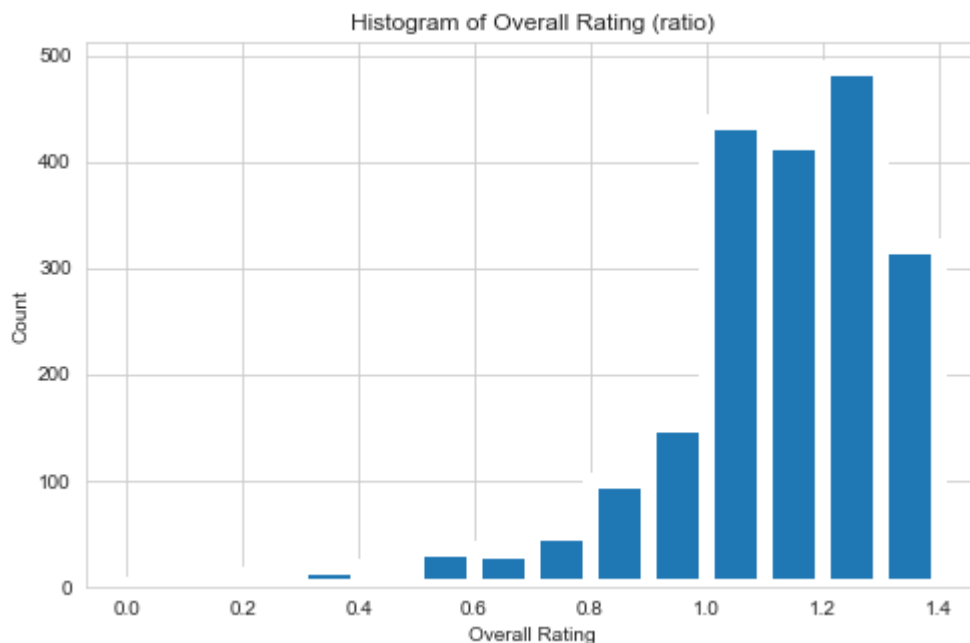


Data Analysis and Visualization

Histogram of Overall Rating

The below histogram shows that the overall rating (ratio of numerator divided denominator) is left skewed. Since there are variations for the numerator and denominator values as discussed in the assess and clean sections, I chose to normalize the rating by taking the ratio of the numerator to denominator. As we can see, the ratings are rather on the higher end and more heavily skewed towards 1.0 and higher. In other words, most ratings are 10/10 or higher.

```
In [251]: # Plot overall rating
plt.figure(figsize = (8, 5))
plt.xlabel('Overall Rating')
plt.ylabel('Count')
bins = [0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1, 1.1, 1.2, 1.3, 1.4]
plt.hist(x = twitter_archive_master.overall_rating,
        edgecolor='white', linewidth=8, bins = bins)
plt.title('Histogram of Overall Rating (ratio)');
```



Dog Stage Pie Chart

From the WeRateDog tweets, it looks like Puppies dominate the 4 categories with 66%, followed by Doggo, Puppo, and Floofer. Keep in mind that not all tweets have a classification on the dog's stage as this is only showing the tweets that do have one.

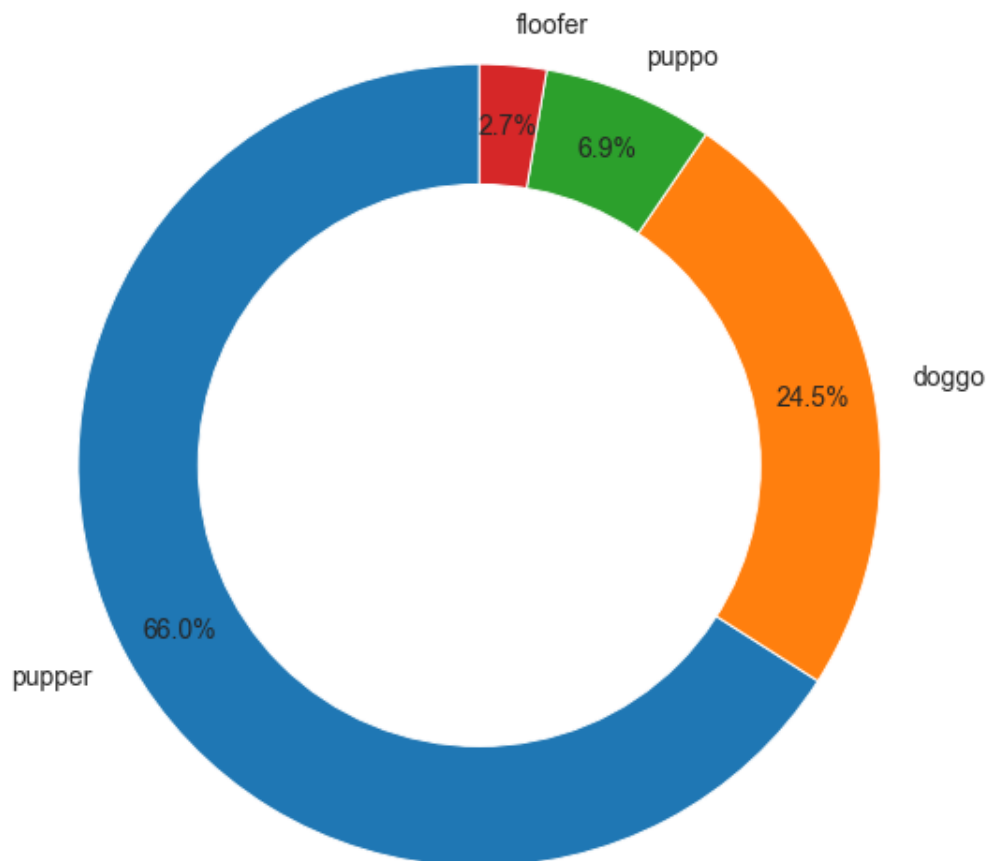
```
In [243]: # Plot the data partitioned by dog stages

stage_count = twitter_archive_master.stage.value_counts().tolist()
stages = twitter_archive_master.stage.value_counts().index.tolist()
fig1, ax1 = plt.subplots(figsize=(9, 7))

ax1.pie(stage_count, labels = stages, autopct='%1.1f%%', shadow = False, start
angle = 90,pctdistance=0.85,
        textprops={'fontsize': 14})
ax1.axis('equal'); # Equal aspect ratio ensures that pie is drawn as a circle.

#draw circle
centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)

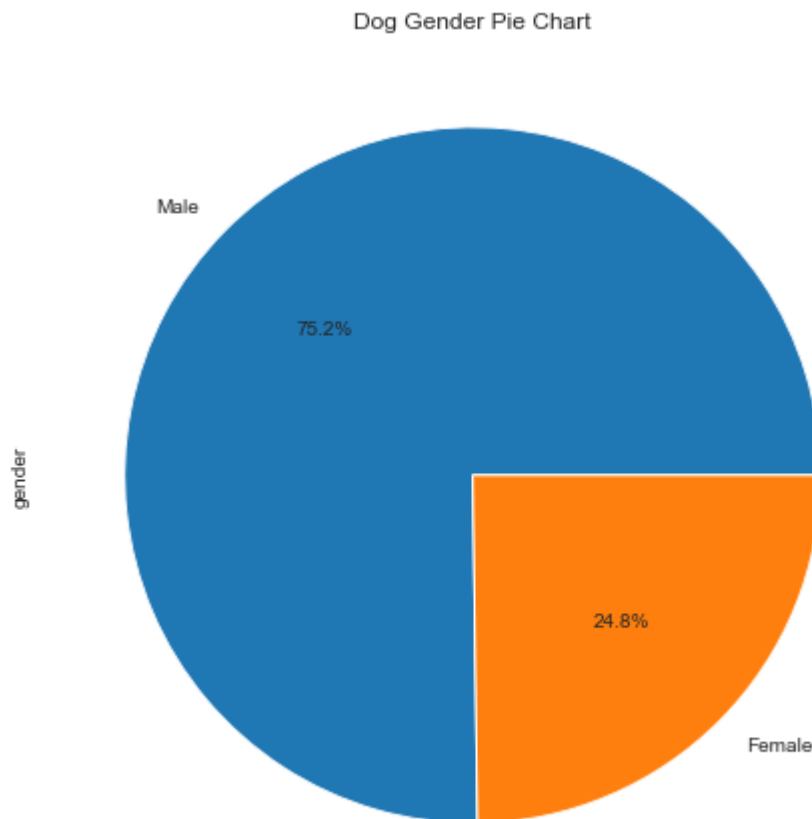
# Equal aspect ratio ensures that pie is drawn as a circle
ax1.axis('equal')
plt.tight_layout()
plt.show;
```



Gender Pie Chart

Here's the dog's gender breakdown from the tweets. Male dogs account for about 75% and female at 25%. The gender is determined by matching the tweet's text to certain keywords. Not all tweets will result in a gender determination.

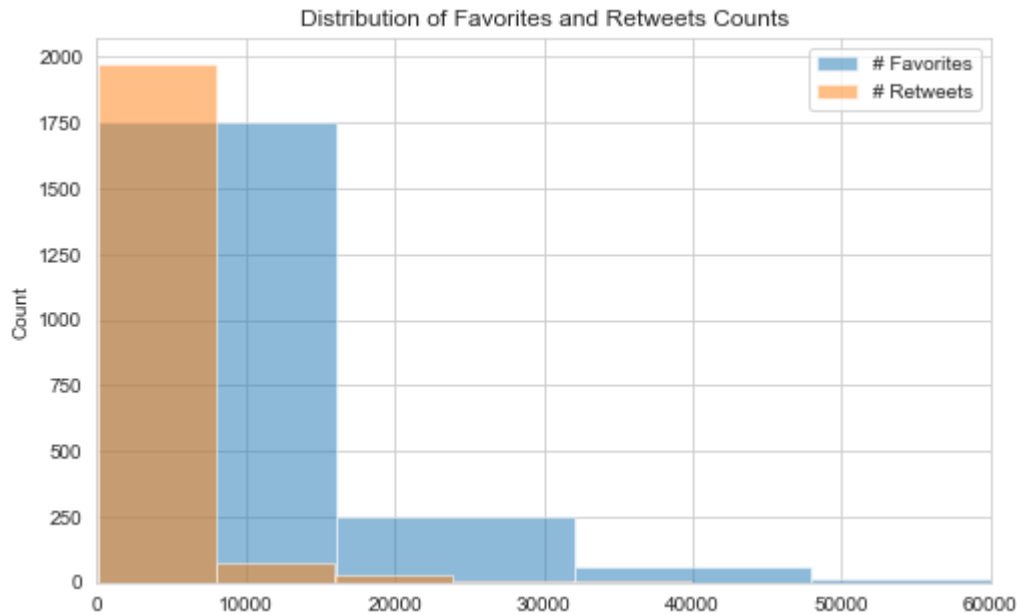
```
In [244]: plt.figure(figsize = (12, 8))  
twitter_archive_master.gender.value_counts().plot(kind = 'pie', autopct='%1.1  
f%%')  
plt.title('Dog Gender Pie Chart');
```



Distribution of Favorites and Retweet Counts

Here's a look at the histogram of the favorite and retweet counts

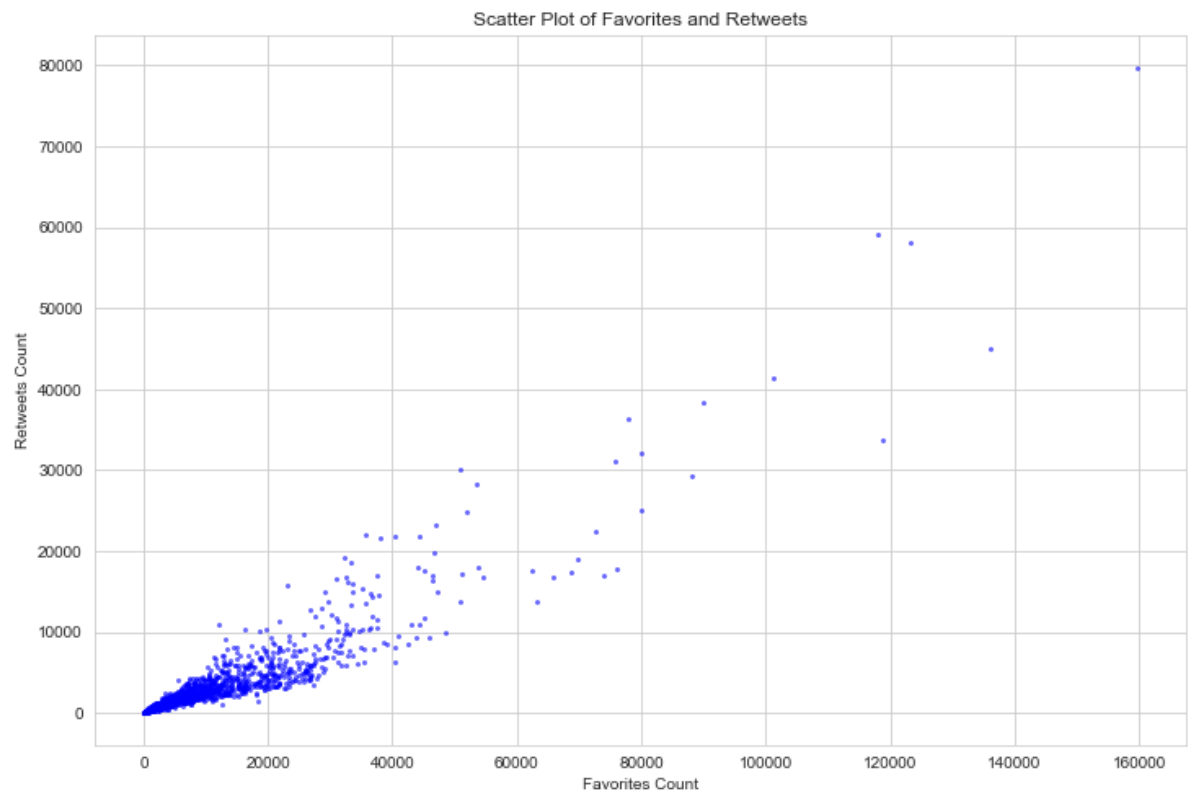
```
In [164]: # Histograms of favorites and retweets
plt.figure(figsize=(8, 5))
plt.xlim(0, 60000)
plt.ylabel('Count')
plt.hist(twitter_archive_master.favorites, alpha=.5, label='# Favorites')
plt.hist(twitter_archive_master.retweets, alpha=.5, label='# Retweets')
plt.title('Distribution of Favorites and Retweets Counts')
plt.legend();
```



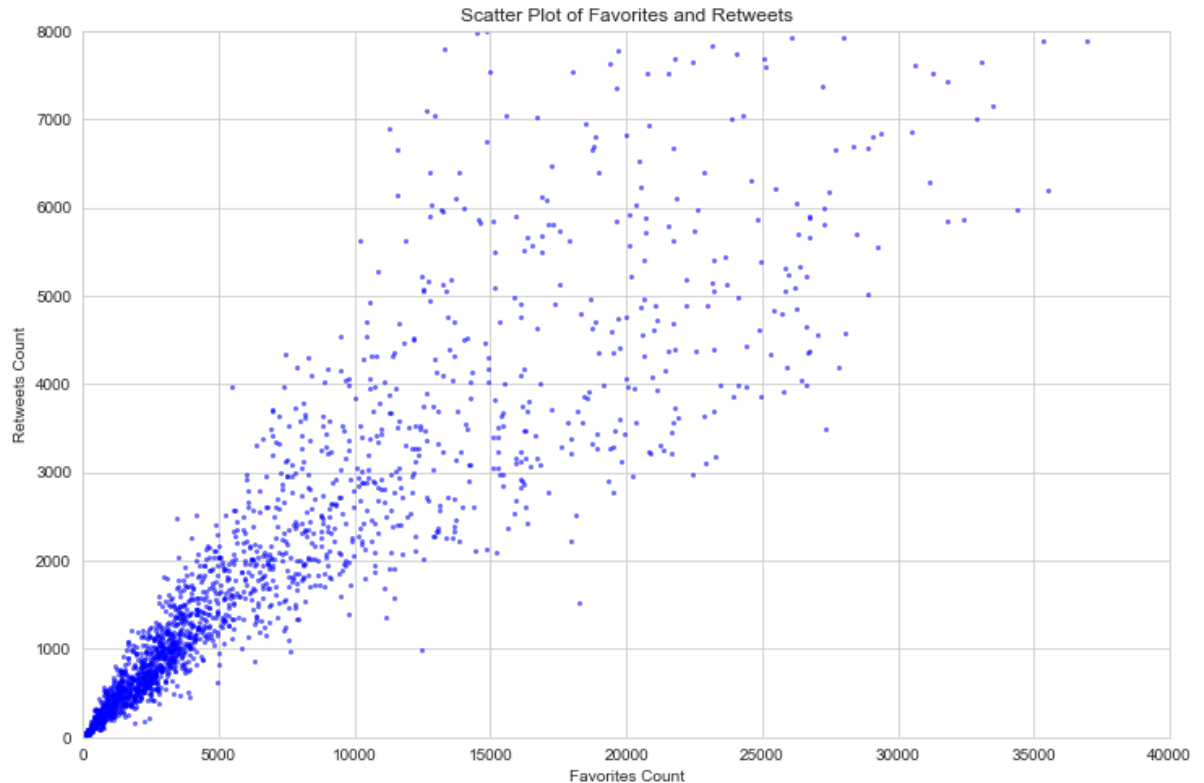
Favorites vs Retweets Scatter Plot

Below are 2 scatterplots of favorite count vs retweet count. From the plot we can see that these 2 variables have a pretty good correlation, as one might expect. The first scatterplot shows all the data points, and I've placed a limits on the x-axis at 4000 and y-axis at 8000 on the second chart to show the points towards the left of the chart in a more magnified view.

```
In [249]: # Scatter plot of favorites and retweets.  
plt.figure(figsize=(12, 8))  
plt.xlabel('Favorites Count')  
plt.ylabel('Retweets Count')  
plt.plot/twitter_archive_master.favorites,  
       twitter_archive_master.retweets, linestyle='', alpha=0.5,  
       marker='o', markersize=2, color='blue')  
plt.title('Scatter Plot of Favorites and Retweets');
```



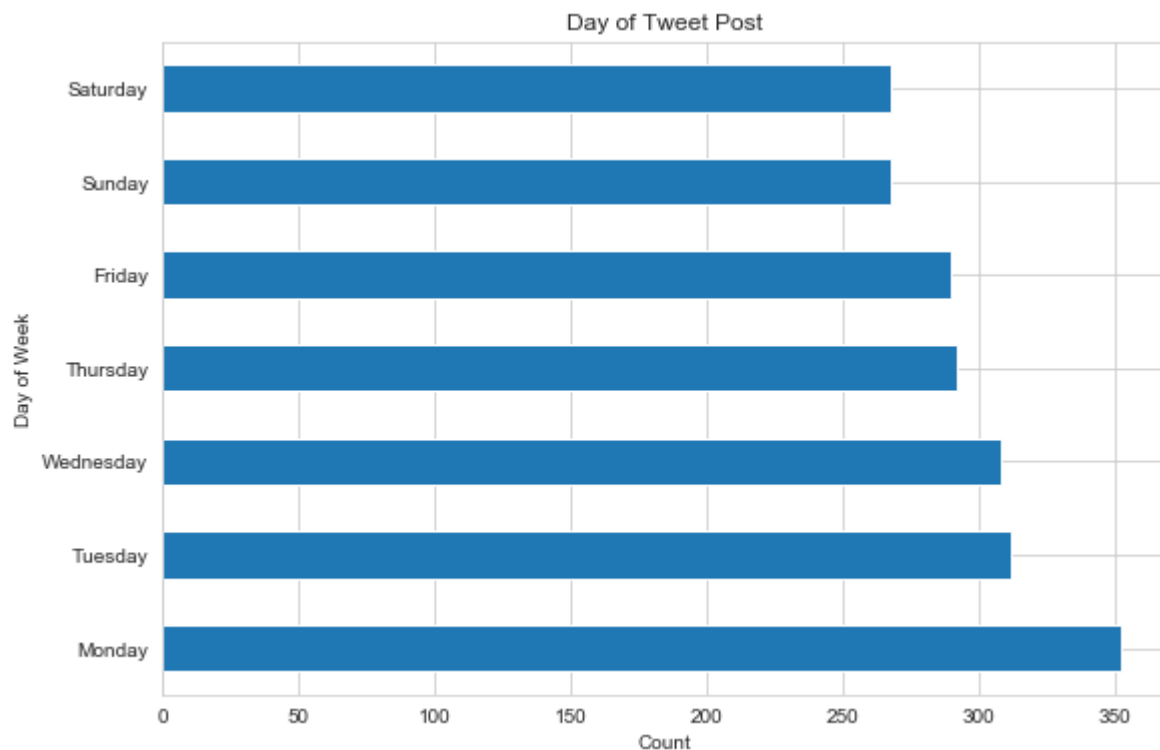
```
In [248]: # Scatter plot of favorites and retweets.
plt.figure(figsize=(12, 8))
plt.xlim(0, 40000)
plt.ylim(0, 8000)
plt.xlabel('Favorites Count')
plt.ylabel('Retweets Count')
plt.plot/twitter_archive_master.favorites,
      twitter_archive_master.retweets, linestyle='', alpha=0.5,
      marker='o', markersize=2, color='blue')
plt.title('Scatter Plot of Favorites and Retweets');
```



Most Popular Day of Week to Post

Interestingly, Monday has the highest count of the day of the week when a tweet is originated based on its timestamp. Maybe people are most stressed on Mondays?

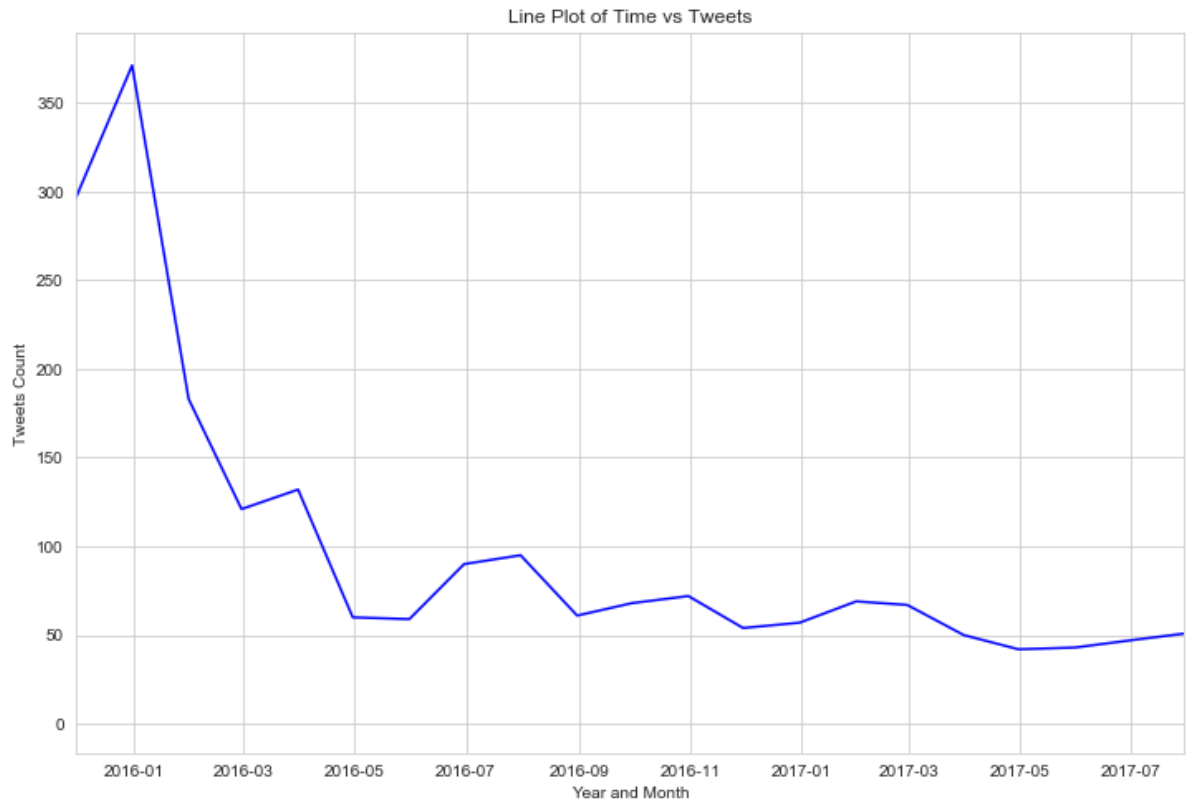
```
In [247]: plt.figure(figsize = (9, 6))
twitter_archive_master.day_of_week.value_counts().plot(kind = 'barh')
plt.title('Day of Tweet Post')
plt.xlabel('Count')
plt.ylabel('Day of Week');
```



Tweet Count by Year and Month

I used the timestamp and grouped them by year and month to determine the number tweets that originated by time period. As we can see, the number of tweets have certainly declined since the peak at January 2016 of over 350 tweets a month and is down to about 50 tweets a month by July 2017.

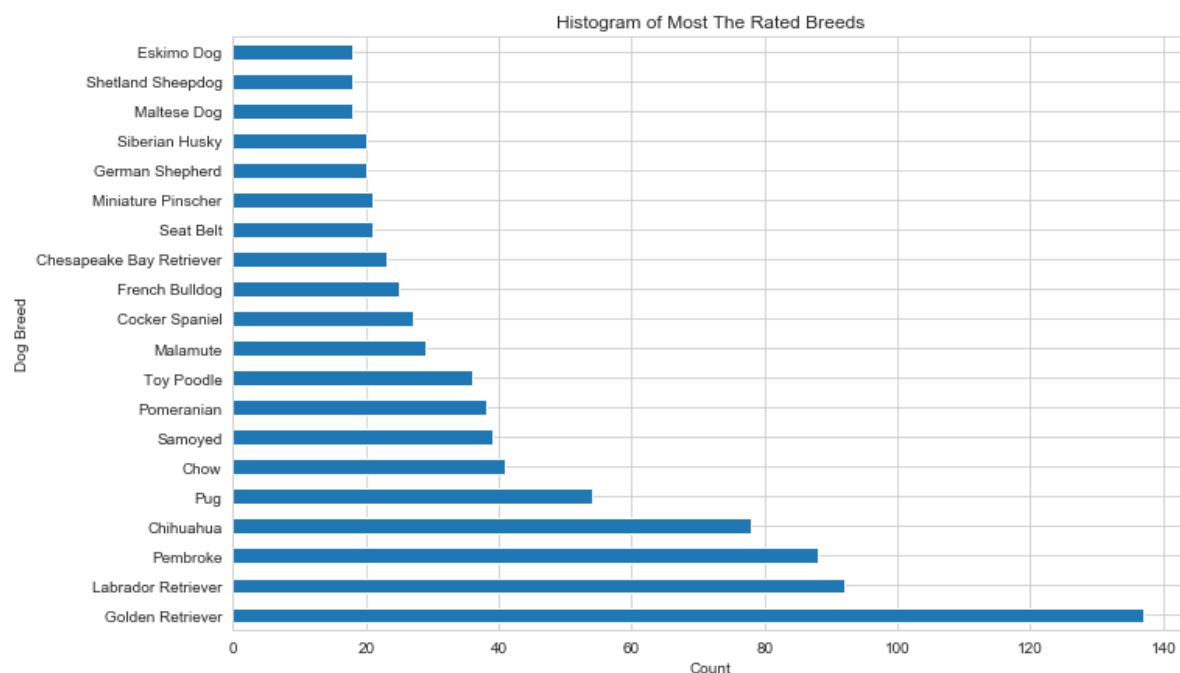
```
In [250]: # Line plot of Time vs Tweets.  
sns.set_style('whitegrid')  
plt.figure(figsize=(12, 8))  
plt.xlim([datetime.date(2015, 11, 30), datetime.date(2017, 7, 30)])  
plt.xlabel('Year and Month')  
plt.ylabel('Tweets Count')  
plt.plot(cg_month_tweets.timestamp, cg_month_tweets.tweet_id, color='blue')  
plt.title('Line Plot of Time vs Tweets');
```



Top 20 Dog Breeds based on tweeted picture

Not suprisingly, Golden Retreivers is most tweeted dog breed in WeRateDogs.


```
In [163]: plt.figure(figsize = (11, 7))
twitter_archive_master.predicted_breed.value_counts().head(20).plot(kind = 'barh')
plt.title('Histogram of Most The Rated Breeds')
plt.xlabel('Count')
plt.ylabel('Dog Breed');
```



Top 20 Dog Breeds based on favorite count

Now I calculate the top 20 dog breeds based on the amount of favorites that the tweet generates. Looks like the Golden Retriever comes out on top again, followed by the beloved Labrador Retriever. These are two of the most popular dogs in the US.

```
In [187]: # Bar graph of most 'favorites' breed.
sns.set_style('whitegrid')
plt.figure(figsize=(12, 8))
plt.barh(range(len(favorite_breed.predicted_breed)), favorite_breed.favorites)
plt.yticks(range(len(fav_breed.predicted_breed)), favorite_breed.predicted_breed)
plt.xlabel('Sum of Favorites')
plt.ylabel('Dog Breed')
plt.title('Bar Chart of The Breeds That Had The Largest Total Number of Favorites');
```

