Visualization

Visualization is the process of finding trends and correlations in data by representing it pictorially.

After gathering, assessing and cleaning data, I merged the three datasets into one and called it “twitter_archive.csv). The first step I took was to read the file in pandas.

```
In [114]: #Reading the csv file in pandas
          master=pd.read_csv('twitter_archive_master.csv')
          master.head()
```

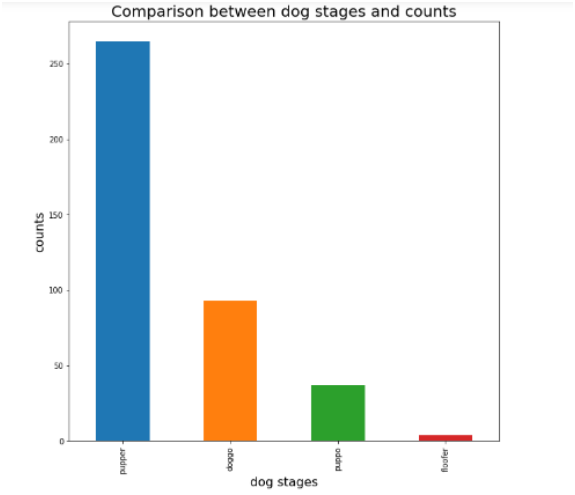
Out[114]:

	Unnamed: 0	tweet_id	timestamp	source	text	expanded_urls	rati
0	0	892420643555336193	2017-08-01	Twitter for iPhone	This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10	https://twitter.com/dog_rates/status/892420643555336193/photo/1	
1	1	892177421306343426	2017-08-01	Twitter for iPhone	This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10	https://twitter.com/dog_rates/status/892177421306343426/photo/1	
				<a	This is Archie. He is a rare Norwegian Pouncing		

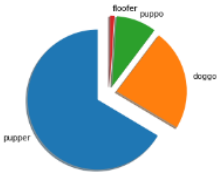
The three insights that I produced was:

1. Comparison in the dog stages by the counts
2. Comparison between retweet count and favorite count
3. Comparison between retweet count and timestamp

For the first insight, it was observed that pupper is the most popular dog stage, followed by doggo, puppo and then floofer being the least. The observation was done on both histogram and a pie-chart. Pupper having a higher turnout compared to the rest would mean that people tend to concentrate more on taking dogs pictures while they are still young.



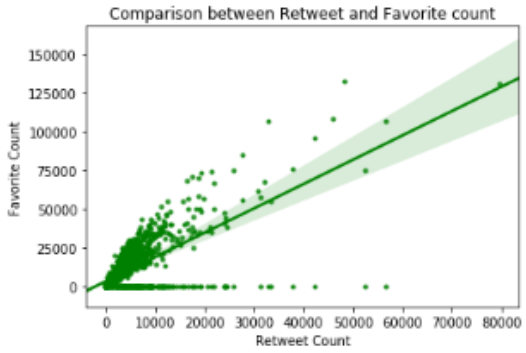
```
In [91]: # Plot the data partitioned by dog stages

dog_stage_count = list(master[master['dog_stage'] != 'None']['dog_stage'].value_counts())[0:4]
dog_stages = master[master['dog_stage'] != 'None']['dog_stage'].value_counts().index.tolist()[0:4]
explode = (0.2, 0.1, 0.1, 0.1)

fig1, ax1 = plt.subplots()
ax1.pie(dog_stage_count, explode = explode, labels = dog_stages, shadow = True, startangle = 90)
ax1.axis('equal')
plt.savefig("data_partitioned.png")
```



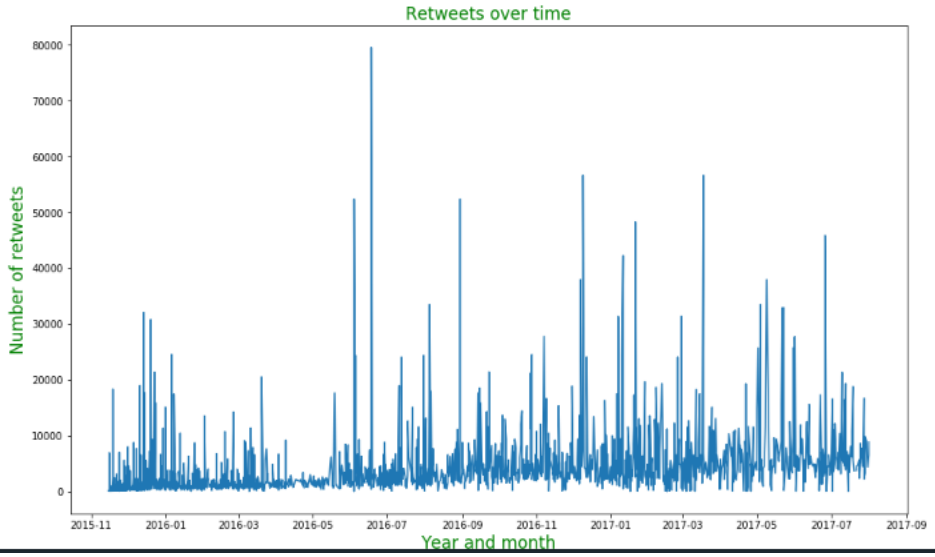
For the second insight, I opted for a scatter plot and then performed a test to get the correlation between retweet and favorite count. The result yielded was +0.70 which indicates a positive correlation between the two variables. That would mean that the higher the number of tweets the higher the possibility for it to be retweeted.

2.Comparison between retweet count and favorite count

```
In [92]: sns.regplot(x= master.retweet_count, y= master.favorite_count, marker= '.', color= 'g'
         plt.xlabel('Retweet Count')
         plt.ylabel('Favorite Count')
         plt.title('Comparison between Retweet and Favorite count');
```



For the third insight, the outcome showed that the number of retweets wasn't consistent over time. In between May and July 2016 there was a higher retweet count compared to other years. The twitter user would use this insight to find out more on what people find interesting so that he can concentrate on that.



In conclusion, many insights would have been drawn out of the dataset, it's just that the project had limits and to dig deeper into the insights would consume a lot of time. A lot of learning has been done in the process, thanks to Udacity.