

# Udacity Data Engineer Nanodegree – Twitter We Rate Dogs Dataset – Act Report – Data Visualization & Insights

My name is Aseem Narula, I am currently working as a Data Engineer at NatWest Group. I have undertaken the Data Engineer Nanodegree.

In this report, I will write down my effort on the “Wrangle and Analyze Data” module in Twitter We Rate Dogs Dataset – Act Report – Data Visualization & Insights.

## Data Visualization & Insights on Wrangled Data

Once tidy master dataset “df\_final\_master\_merged” gathered, assessed, and cleaned master dataset(s) to a CSV file with the name - "twitter\_archive\_master.csv".

Now it's time to present the insights on the cleaned wrangled data.

Now, our data is ready for the data visualization and let's find some interesting insights, I will try to answer the following questions through this.

**Insights #1** — What are the top 3 dog breed types with the favourite count?

### Insights #1

```
# What are the top 3 dog breed types with the favourite count ?  
|  
df_final_master_merged.sort_values('favorite_count', ascending=False)[['dog_breed_stage', 'favorite_count']].head(3)
```

	dog_breed_stage	favorite_count
51	puppo	132810
134	doggo	131075
18	pupper	105827

*The dog breed stage - "puppo" is the most favourite among all i.e. there are total of the 132810 favourite count tweets.*

*The dog breed stage - "puppo" is the most favourite among all i.e. there are total of the 132810 favourite count tweets.*

## Insights #2 — Which dog breed stage is most common in our twitter dataset?

### Insights #2

```
# Which dog breed stage is most common in our twitter dataset :|  
df__dog_breed_stage_by_tweet_id = df_final_master_merged.groupby('dog_breed_stage')['tweet_id'].count()
```

```
df__dog_breed_stage_by_tweet_id
```

```
dog_breed_stage  
doggo          63  
doggofloofer    1  
doggopupper     9  
doggopuppo      1  
floofer         7  
pupper        203  
puppo          22  
Name: tweet_id, dtype: int64
```

*The dog breed stage -"pupper" is the most comon dog breed stage among all i.e. there are total of the 203 which means the most common dog breeds which are seen is "pupper".*

*The dog breed stage -"pupper" is the most comon dog breed stage among all i.e. there are total of the 203 which means the most common dog breeds which are seen is "pupper".*

**Insights #3** — What are the most top 3 dog breed stages with maximum retweet count number ?

**Insights #3**

```
# What are the most top 3 dog breed stages with maximum retweet count number ?
df_final_master_merged.sort_values('retweet_count', ascending=False)[['dog_breed_stage', 'retweet_count']].max()
dog_breed_stage    puppo
retweet_count      79515
dtype: object
```

*The dog breed stage - "puppo" is the most retweeted stage with the maximum of the 79515 retweets.*

**The dog breed stage — “puppo” is the most retweeted stage with the maximum of the 79515 retweets.**

**Insights #4** — How are tweets are from each different sources ?

**Insights #4**

```
: # How are tweets are from each different sources ?
| df_final_master_merged.groupby('source_x')['tweet_id', 'source_x',].count()
:
```

	tweet_id	source_x
source_x		
TweetDeck	2	2
Twitter Web Client	1	1
Twitter for iPhone	303	303

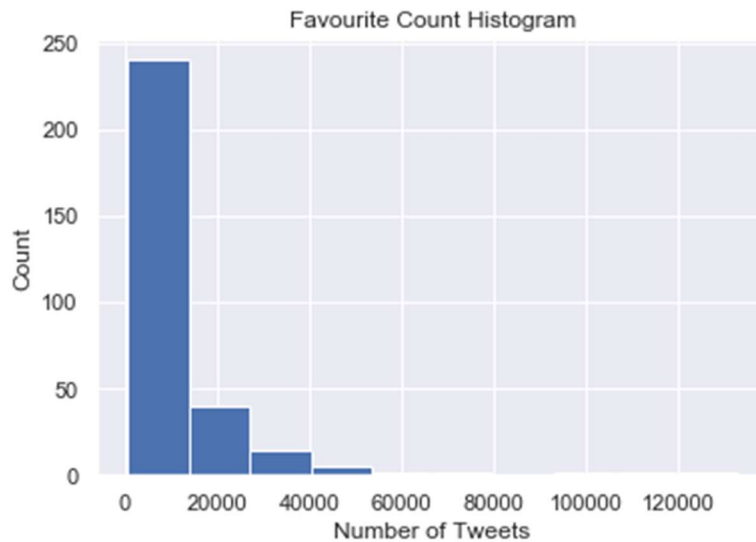
*There are maximum number of the tweets from the "Twitter for iPhone" in the final clean merged dataset.*

**There are maximum number of the tweets from the “Twitter for iPhone” in the final clean merged dataset.**

## Data Visualization

Plotting histogram for the favourite count column variable- The number of tweets which are marked as “favourite tweets” are maximized in the range of the 200–250 where are least favourite tweets are in the range of the 40000–60000.

```
fig, ax1 = plt.subplots()
plt.hist(df_final_master_merged.favorite_count)
plt.title('Favourite Count Histogram');
ax1.set_ylabel('Count');
ax1.set_xlabel('Number of Tweets');
```



Checking the correlation between the Rating Numerator and the Retweet Count, We can see that there is direct correlation between the rating numerator and retweet count, the tweet/post which are having higher rating numerator are mostly retweet in the range starting from the 60000 to 80000.

```
sns.regplot(df_final_master_merged.retweet_count, df_final_master_merged.rating_numerator);
plt.title('Correlation between Rating Numerator and Retweet Count');
```

