Aseem Narula

6 min read · Just now

Udacity Data Engineer Nanodegree — Twitter We Rate Dogs Dataset Act Report

Introduction

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

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "**they're good dogs Brent**." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Project Overview Steps

Tasks in this project are as follows:

Step 1: Gathering data
Step 2: Assessing data
Step 3: Cleaning data
Step 4: Storing data
Step 5: Analyzing, and visualizing data
Step 6: Reporting — Data wrangling efforts &Data Visualization with Insights

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

I will breakdown as per the project rubrics specifications-

Code Functionality and Readability

The Python code written in the Juypter Notebook- wrangle_act.ipynb

· Important necessary libraries are imported in the beginning of the Jupyter Notebook

· Code are prefixed with the comment lines giving the brief description what their functionality is all about.

Gathering Data

In this step, the We Rate Dogs Twitter data is gather from the following sources

· **Twitter Archive Enhanced Dataset CSV file**

· This file is downloaded from the Udacity Server and manually uploaded to the local path where the Juypter Notebook is placed.

· **Images Predictions TSV File**

· This file is downloaded from the Udacity Server programmatically using the Requests library and manually uploaded to the local path where the Juypter Notebook is placed.

· **Twitter Additional Data File JSON File**

· This file was supposed to the downloaded via Twitter API using Tweepy library but code failed while Twitter API settings authentication failed and so I had to raise ticket with Knowledge support team, Mentor has advised me to use the ‘tweet-json.txt’ dataset and continue the module

· ### Ticket Reference — https://knowledge.udacity.com/questions/986996

I have loaded the 3 above source files into the three separate dataframe in Python using Pandas library.

1. df_twitter_archive_enhanced_dataset

2. df_twitter_image_predictions_tsv

3. df_tweet_json_file

Note — I have continued the further into my local PC environment where Python Anaconda package is installed because the Udacity cloud workspace was getting disconnected frequently.

Assessing Data

Once the data is gathered and loaded in the respective dataframe, I have analyzed and assessed the data using the following functions for each of the data frames.

Info, Describe, Shape, value_counts, duplicated, isnull, isna

I have found the following issues

Twitter Archive Enhanced Dataset

1. Timestamp field is string object.
2. The data columns — tweet_id,in_reply_to_status_id,in_reply_to_user_id etc. should be an object data type but currently they of data types — integer and float in the original dataset.
3. There are only 181 retweets — columns — retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.We need only original “tweets” not “retweets”.
4. There are “in_reply_to_status_id” and “in_reply_to_user_id” columns which are extra as we need only “original tweets” column in our dataset.
5. There are 109 inconsistent names with small letters with no valid names,also there are NaN values in the dataset in the “name” column should be rename to ‘None’.
6. There is one record with rating denominator with the value Zero.
7. There are rating denominator values greater than 10, should be removed.
8. There are only 4 types of values in the source column which can be simplified further to remove complexity. Twitter for iPhone, Vine — Make a Scene, Twitter Web Client, TweetDeck

9. There are 59 number of tweets which are having missing expanded URL's

Twitter Image Prediction Dataset

1. There are 324 rows in the Twitter Image Prediction Dataset dataframe which didn't predict anything for three of dog type, best way to predict the dog breed is to combine it with master dataset.

Tidiness issues

1. The 4 Columns doggo, floofer, pupper, and puppo — should be merged into the one column as this doesn't comply "tidy data" rule.
2. Json Tweet text file dataset should be union with 'Twitter Enhanced Archive' Dataset to get the full view.
3. The data columns in the Json Tweet text file dataset should be reduced to the "tweet_id", "recount_tweet", "favourite_count", extra columns must be removed.

Cleaning Data

Copies of the original pieces of data are made prior to cleaning using df.copy() function

a) df_twitter_archive_enhanced_dataset_copy

b) df_twitter_image_predictions_tsv_copy

c) df_tweet_json_file_copy

All issues identified in the assess phase are successfully cleaned (if possible) using Python and pandas, and include the cleaning tasks required.

a) Merging columns — "doggo", "floofer","pupper","puppo" into single column called — "dog_breed_stage"

b) Dropping the columns — 'doggo', 'floofer', 'pupper', 'puppo' which are no longer required

c) Checking the dataset to confirm if columns are clubbed properly.

d) Merging the Twitter Enhaced dataset with the Image Predictions dataset after cleaning

e) Renaming the "id" column to the "tweet_id" in the tweet json file dataset so that it can be

further used for the merge in the next step

A tidy master dataset “**df_final_master_merged**“ with all pieces of gathered data is created.

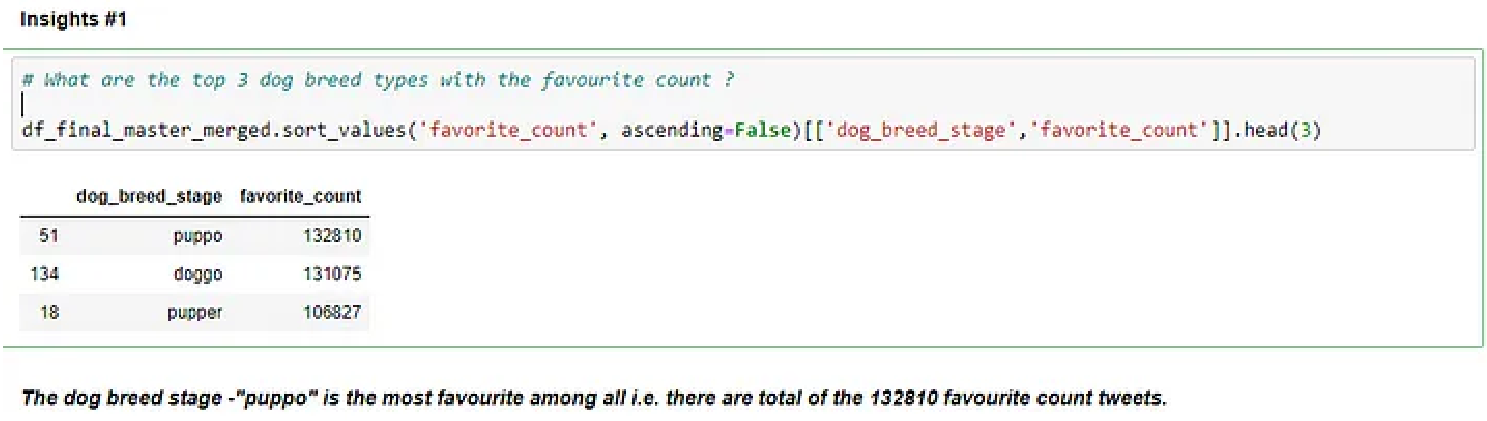
Storing and Acting on Wrangled Data

A 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”.**

Data Visualization and Insights

Now, our data is ready for the data visualization and lets 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

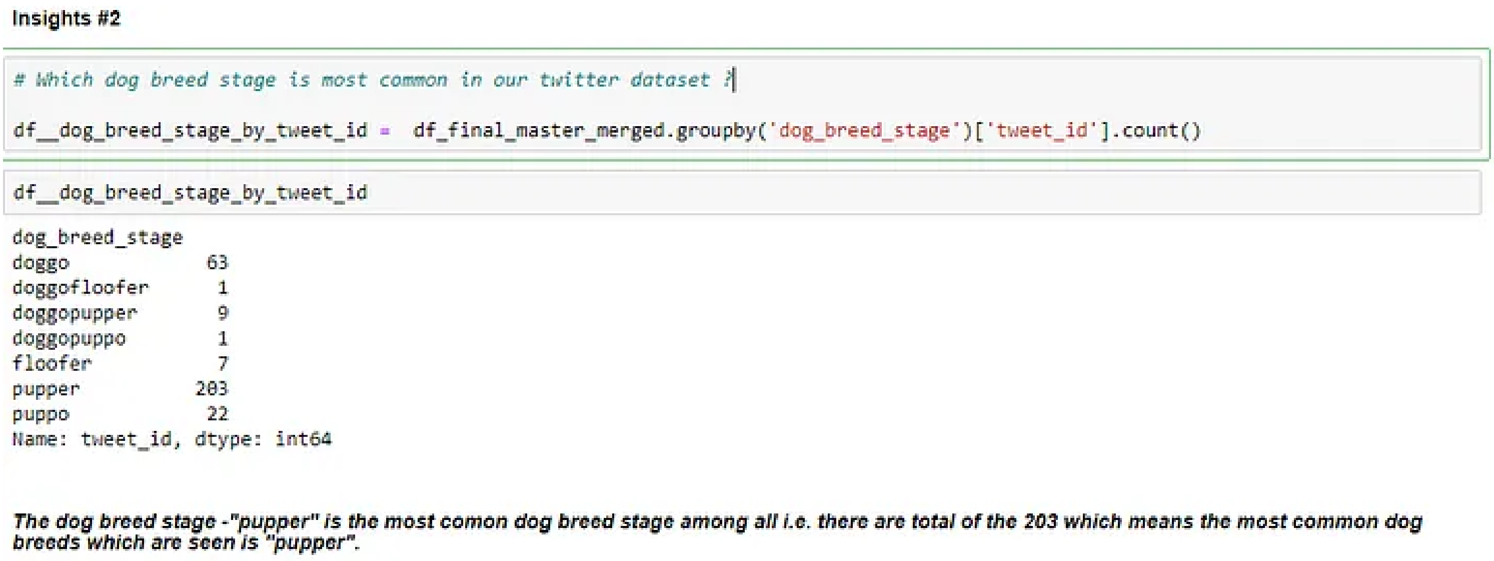
```
# What are the top 3 dog breed types with the favourite count ?

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	106827

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()
```

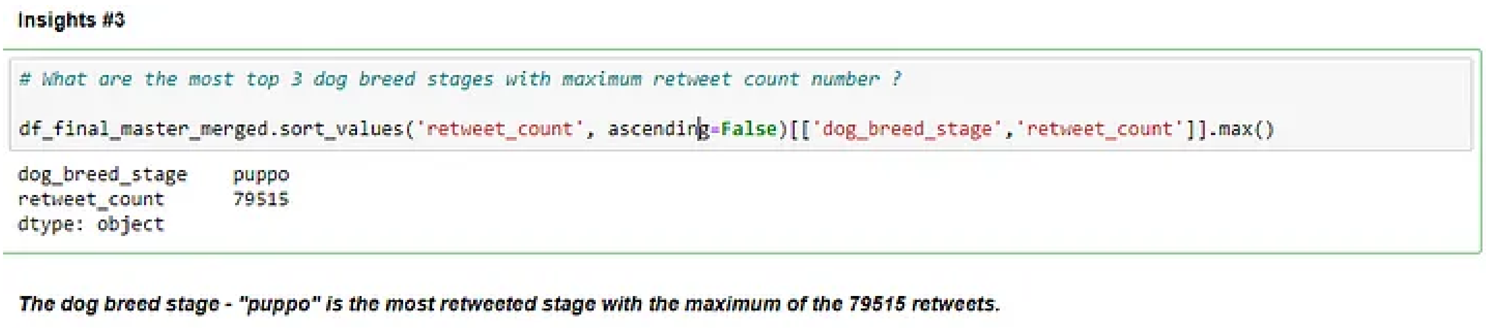
```
df__dog_breed_stage_by_tweet_id
```

```
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

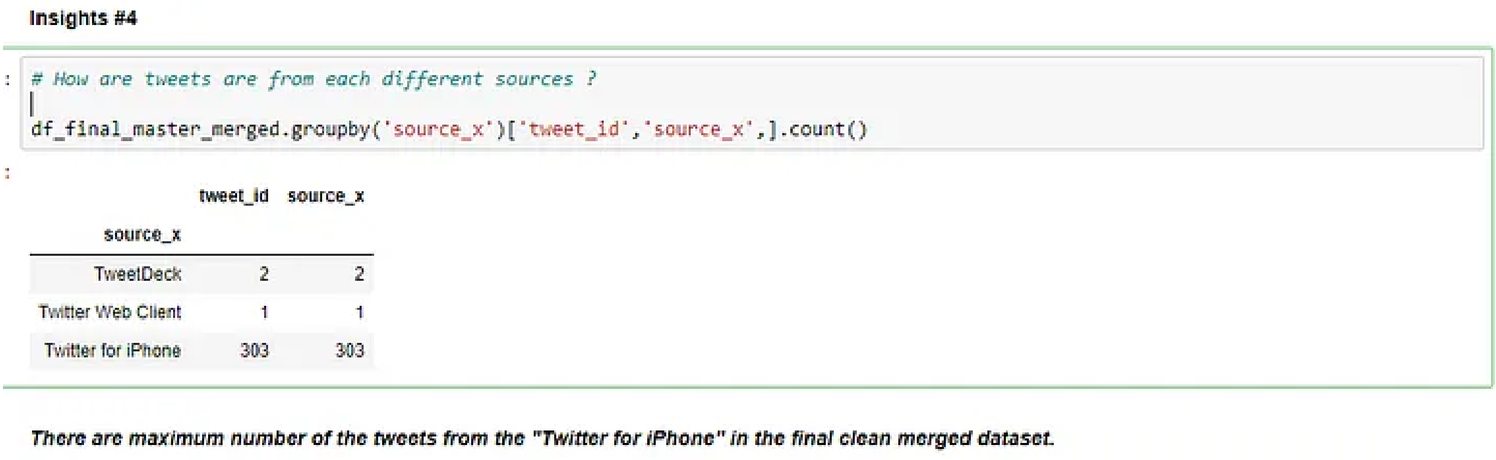
```
# What are the most top 3 dog breed stages with maximum retweet count number ?

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()
```

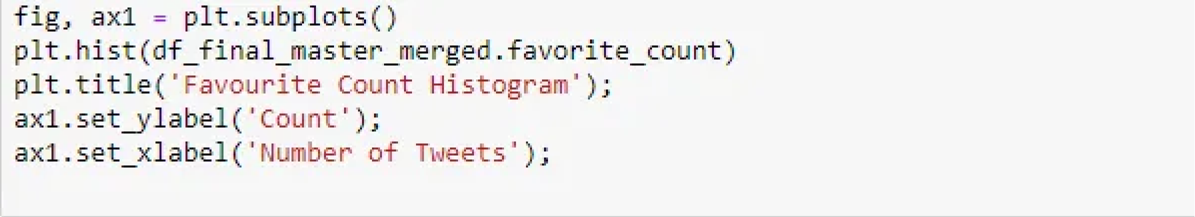
	tweet_id	source_x
source_x		
TweetDeck	2	2
Twitter Web Client	1	1
Twitter for iPhone	303	303

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

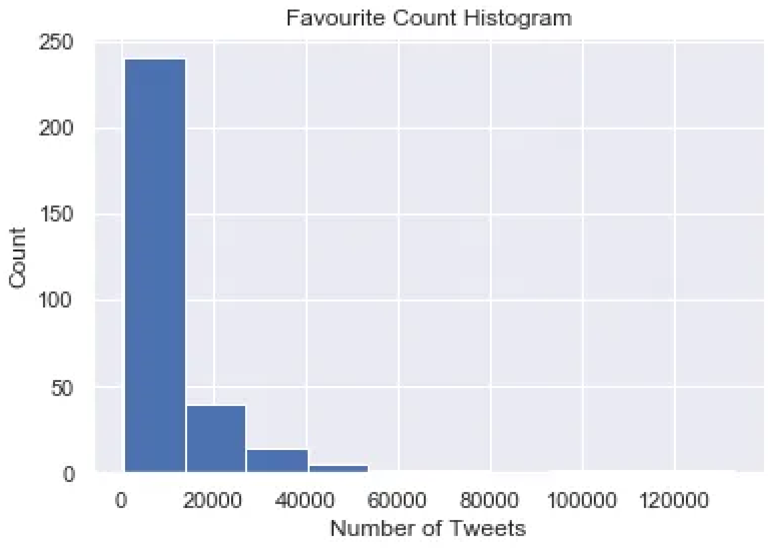
There are maximum number of the tweets from the "Twitter for iPhone" in the final clean merged dataset.

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.

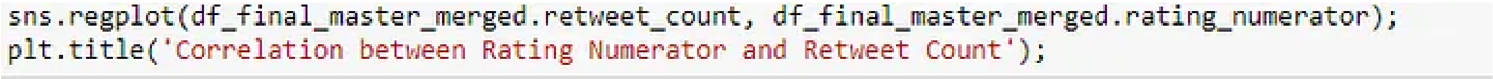


```
fig, ax1 = plt.subplots()
plt.hist(df_final_master_merged.favorite_count)
plt.title('Favourite Count Histogram');
ax1.set_ylabel('Count');
ax1.set_xlabel('Number of Tweets');
```

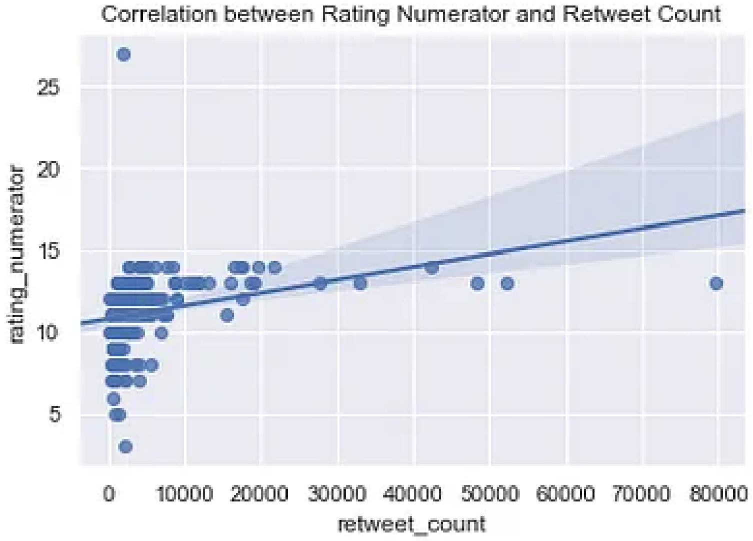


# Checking the correlation between the Rating Numerator and the Retweet Count

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');
```



Acknowledgement

All the datasets of We Rate Dogs Twitter account used in this Data Engineer Project are provided through Udacity and are used for my project with Udacity Data Engineer Nanodegree.