WeRateDogs Udacity Project

Wrangling Efforts

"A brief report to outline the wrangling efforts, type of data and analysis, and the insights extracted from twitter WeRateDogs data archives"

1. **Data Gathering:**

We had an extracted file already contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017, however it wasn't enough to generate useful insights so we had to augment the data with two additional files.

First i downloaded from the twitter_archive.csv manually, then image prediction.tsv was downloaded **using Requests/OS libraries**, Lastly the twitter API was downloaded using **tweepy** library (the method I followed was for a video on YouTube applying this function to search for hashtags, I had the idea and applied it I find it very useful if I am working with a team to allow other coders to use their credentials to have the most updated json, then I assigned the required headers for assessing, cleaning, and generating insights.

Converted all three types to pandas **DataFrames** and moved to the next phase.

2. **Data Assessing:**

I Have made both Visual and Programmatic assessment, and some of the visual have been done using google sheets, and here are screenshots for (Quality Issues) Detected while visual assessment.

tweet_id

A	B	C	D	E	F
tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text
674339060434379000	665814696700723000	16374678	2015-12-08 20:53:11 +0000	Twitter for iPhone	13/10 @ABC7
667070482143944000	667066535570550000	4196983835	2015-11-18 20:02:51 +0000	Twitter for iPhone	After much debate this dog is being upgrac
863079547188785000	667152164079423000	4196983835	2017-05-12 17:12:53 +0000	Twitter for iPhone	Ladies and gentlemen... I found Pipsy. He may have changed his name to Pablo, but he nev
704871453724954000	667152164079423000	4196983835	2016-03-02 03:30:25 +0000	Twitter for iPhone	I found a forest Pipsy. 12/10 https:/
669353438988365000	667806454573760000	4196983835	2015-11-25 03:14:30 +0000	Twitter for iPhone	This is Tessa. She is also very pleased after finally meeting her
668967877119254000	668920717132582000	21435658	2015-11-24 01:42:25 +0000	Twitter for iPhone	12/10 good shit Bu @wane15
669684665554620000	669354382627049000	4196983835	2015-11-26 01:11:28 +0000	Twitter for iPhone	After countless hours of research and hundreds of formula alterations we
694366675654983000	670668383497736000	4196983835	2016-02-02 03:08:26 +0000	Twitter for iPhone	This pupper only appears through the hole of a Funyun. Much like Phineas, t
671550332464455000	671544874165002000	4196983835	2015-12-01 04:44:10 +0000	Twitter for iPhone	After 22 minutes of careful deliberation this dog is being demoted to a 1/10
671729906628341000	671561002136281000	4196983835	2015-12-01 16:37:44 +0000	Twitter for iPhone	I'm just going to leave this one here as well. 1
674793399141146000	671729906628341000	4196983835	2015-12-10 03:30:58 +0000	Twitter for iPhone	I have found another. 13/10 https:/
673716207231690000	673715861853720000	4196983835	2015-12-07 04:11:02 +0000	Twitter for iPhone	The millennials have spoken and we've decided to imr
674606911342424000	674468880899788000	4196983835	2015-12-09 15:09:55 +0000	Twitter for iPhone	The 13/10 also takes into account this impeccable yard. Louis is great but
674742531037511000	674739953134403000	4196983835	2015-12-10 00:08:50 +0000	Twitter for iPhone	Some clarification is required. The dog is singing Cher and th
674754018082705000	674752233200820000	4196983835	2015-12-10 00:54:28 +0000	Twitter for iPhone	Just received another perfect photo of dogs and the s
674999007681908000	674793399141146000	4196983835	2015-12-10 17:11:09 +0000	Twitter for iPhone	Ok last one of these. I may try to make some myself. Anywa
675349384339542000	674999007681908000	4196983835	2015-12-11 16:20:15 +0000	Twitter for iPhone	Yea I lied. Here's more. All 13/10 http
675676694217588000	675349384339542000	4196983835	2015-12-06 00:35:13 +0000	Twitter for iPhone	If you are aware of who is making these please let me know:
675707330206547000	675497103322386000	4196983835	2015-12-12 16:02:36 +0000	Twitter for iPhone	We've got ourselves a battle here. Watch out Re
675870721063669000	675707330206548000	4196983835	2015-12-13 02:51:51 +0000	Twitter for iPhone	& this is Yoshi. Another world record contender 11/10 (what the hell is happen
675849018447167000	675845657354215000	4196983835	2015-12-13 01:25:37 +0000	Twitter for iPhone	This dog is being demoted to a 9/10 for not wearing a helmet w
676590572941893000	676588346097852000	4196983835	2015-12-15 02:32:17 +0000	Twitter for iPhone	After some outrage from the crowd. Bubbles is being upgraded t
678023323247357000	678021115718029000	4196983835	2015-12-19 01:25:31 +0000	Twitter for iPhone	After getting lost in Reese's eyes for several minutes
681340665377193000	681339448655802000	4196983835	2015-12-28 05:07:27 +0000	Twitter for iPhone	I've been told there's a slight possibility he's checking his r
682808988178739000	682788443153760000	4196983835	2016-01-01 06:22:03 +0000	Twitter for iPhone	I'm aware that I could've said 20/16, but here at WeRateDogs we are very pr
684225744407494000	684222868335505000	4196983835	2016-01-05 04:11:44 +0000	Twitter for iPhone	Two sneaky puppers were not initially seen, moving the rating to 143/13
684538444857667000	684481074559381000	4196983835	2016-01-06 00:54:18 +0000	Twitter for iPhone	After watching this video, we've determined that Pippa will be upgrac
684969868808454000	684959798585110000	4196983835	2016-01-07 05:28:35 +0000	Twitter for iPhone	For those who claim this is a goat, u are wrong. It is not the Greatest Of All T
685681090388975000	685547936038666000	4196983835	2016-01-09 04:34:45 +0000	Twitter for iPhone	Jack deserves another round of applause. If you missed this earlier today
686035780142297000	686034024800862000	4196983835	2016-01-10 04:04:10 +0000	Twitter for iPhone	Yes I do realize a rating of 4/20 would've been fitting. However, it would b
690607260360429000	690341253549002000	467036706	2016-01-22 18:49:36 +0000	Twitter for iPhone	12/10 @Lightningh
746818907684614000	691416866452082000	4196983835	2016-06-25 21:34:37 +0000	Twitter for iPhone	Guys... Dog Jesus 13/10 buoyant af https://t.co/

in_reply_to_status_id	Int64	*Nullable.* If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID. Example: `"in_reply_to_status_id":1051222721923756032`
in_reply_to_status_id_str	String	*Nullable.* If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's ID. Example: `"in_reply_to_status_id_str":"1051222721923756032"`

I also assessed the two other files, (status.json, and image_prediction.tsv) visually and programmatically in jupyter notebook using methods such as .info(), .head(), .sample(), and .value_counts().

The datasets were accessed to scan two criteria, quality and tidiness. When an issue was detected it was documented under one of these twocriteria based on the issues found.

3. **Data Issues:**

Tidiness Issues

After assessing the data provided here is a summary of the found issues.

There are 4 columns (doggo,floofer, pupper, puppo) which are dog stages in DoggoLingo and the info provided in the project motivation, means there is one variable stored in 4 columns, Some columns doesn't represent any values ('expanded_urls', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'), and might not be needed for analysis, NONE word needs to be replaced to allow us to know the null values, Timestamp has more than one variable (Date, and Time), (p1, p1_conf, p1_dog) Don't represent variable names, The three data sets has columns that needs to be removed, Not useful to analyze or capture insights, and lastlyThe three data sets are same observation unit displayed in 3 tables.

Quality Issues:

Columns with missing values [in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls] (Completeness Issue).

Based on the requirement it has to be original tweet neither retweets nor reply, tweets that has value in_reply_to_status_id, needs to be removed and keep original tweets only (Validity Issue), rating_numerator column has 21 outliers, rates above the accepted range (from 44 to 1776)(validity & accuracy), "These ratings almost always have a denominator of 10".said in the (Project Overview), however we have values in rating_denominator columns less and more than 10, 'rating_denumirator' has zero values, 'Name' column has too many NONE values, might cause issues analyzing, Timestamp needs to be converted to PD.datetime values, 66 Duplicate values in column jpg_url, and more than 2000 duplicated (img_num), P1 column has prediction of no dogs names, Columns from p2 until p3_dog have low condidence levels, needs to be removed, we have in P1_dog 543 values for no dogs (False), And API dataset has 163 values favorite_count == zero, means never been favorited

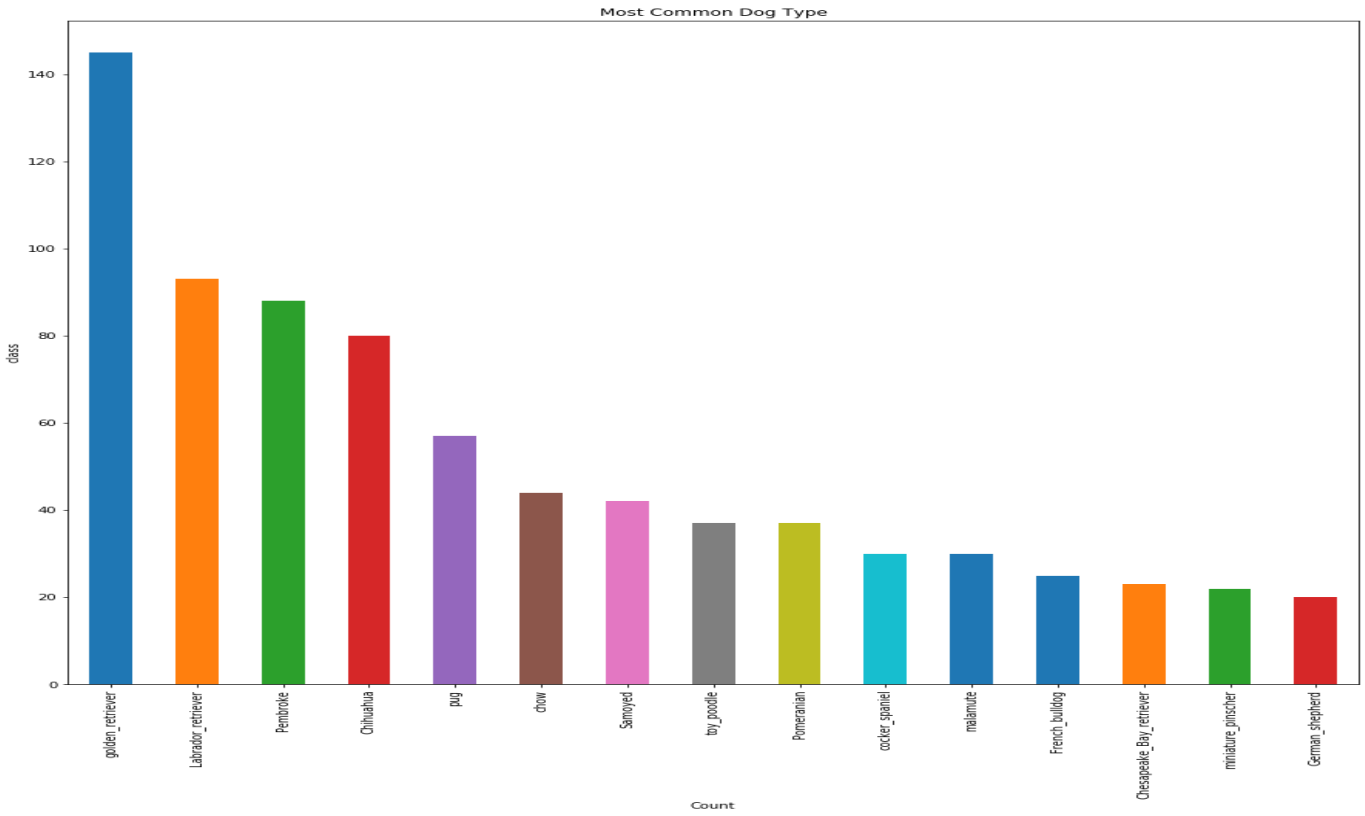
Data Cleaning:

The final step in the wrangling process is cleaning the data for quality and tidiness issues, first I created a cleaned data sets by copying, remove all rows that has values (in_reply_to_status_id), Remove Columns with Missing values, and excluded from the analysis, Remove all the outliars in column (rating_numerator), Cleaning the column rating_denumirator from all values != 10, Etc until created one master Dataframe also merged the four dog stage columns into one, ETC (**all explained in the notebook as well).**

- **Insights and visuals:**

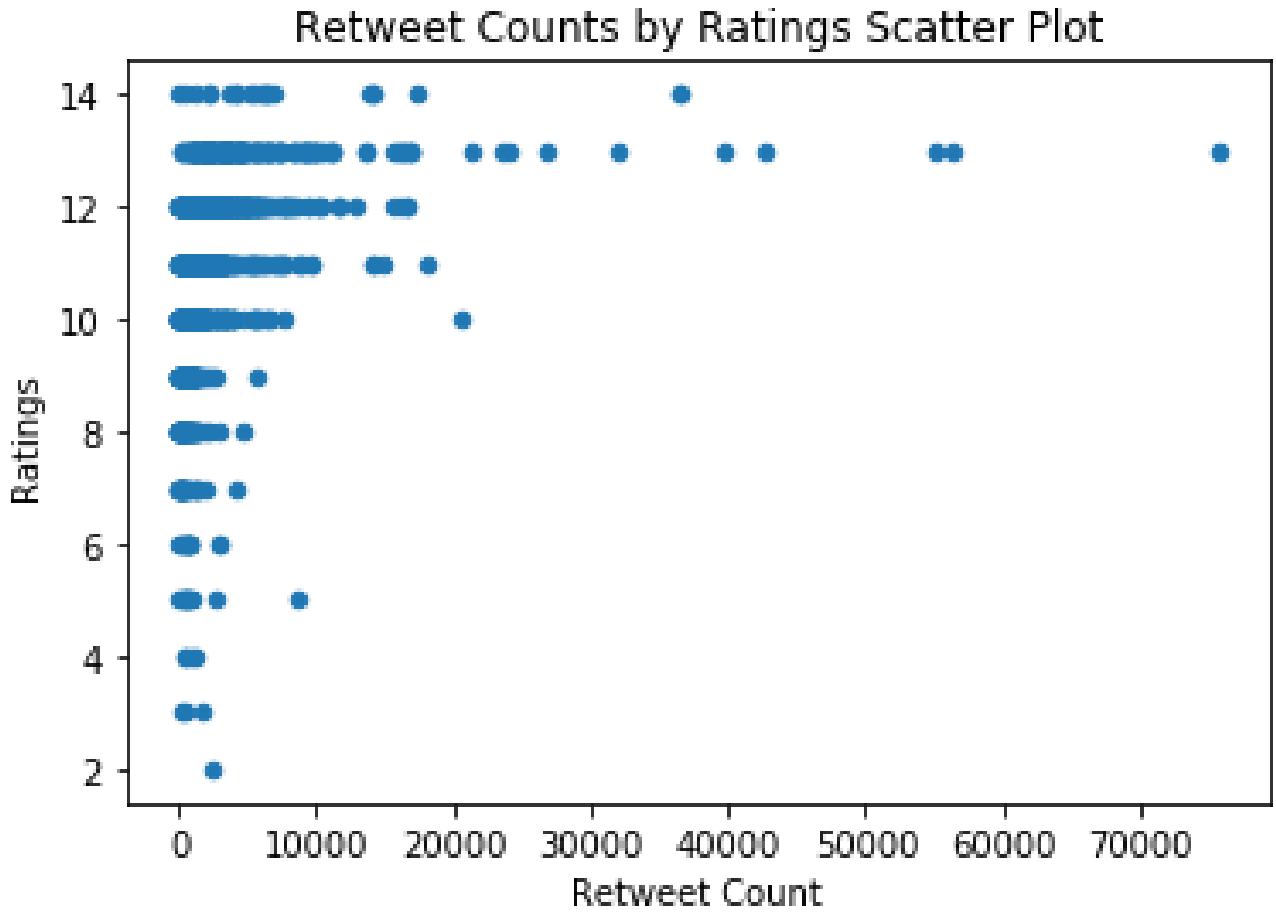
1. **First Insight:**

"golden_retriever, and Labrador_retriever are the most common dog according to WeRateDogs Data, and Aim and West_Highland_white_terrier is the less common dog ccording to WeRateDogs Data, and AI"



2. Second Insight:

"Top Rated Dogs have less retweet count, and vice versa"



3. **Third Insight:**

"The top rated dog class is not the most common class. The most rated dog is Saluki, Tibetan_mastiff, briard, Border_terrier, standard_schnauzer, silky_terrier"

4. **Forth Insight:**

"Dog ratings has no correlation with the most common dog"