Introduction

In this project we will follow the complete data analysis process (Gather, Asses, Wrangle, analyze, and visualize).

The dataset that we will be wrangling is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gather

There are three data sources for this project we need to gather :

1- **CSV Data File**: "twitter-archive-enhanced.csv" there are 2356 records which is a twitter archive basic tweets data , including rating, dog name, and dog "stag:e" (i.e. doggo, floofer, pupper, and puppo) . data loaded from this source into **df_csv** data frame with below structure :

#	Column	Non-Null Count	Dtype
0	tweet_id	2356 non-null	int64
1	in_reply_to_status_id	78 non-null	float64
2	in_reply_to_user_id	78 non-null	float64
3	timestamp	2356 non-null	object
4	source	2356 non-null	object
5	text	2356 non-null	object
6	retweeted_status_id	181 non-null	float64
7	retweeted_status_user_id	181 non-null	float64
8	retweeted_status_timestamp	181 non-null	object
9	expanded_urls	2297 non-null	object
10	rating_numerator	2356 non-null	int64
11	rating_denominator	2356 non-null	int64
12	name	2356 non-null	object
13	doggo	2356 non-null	object
14	floofer	2356 non-null	object
15	pupper	2356 non-null	object
16	puppo	2356 non-null	object

2- **Twitter API Data:** additional json formatted data can be gathered by Calling Twitter's API to query the twitters using tweetID . Code to call twitter API is provided in "*Twitter-api.py"* and data gathered is saved in "*tweet_json.txt*" . then loaded from this file into **df_json** data frame with below structure :

#	Column	Non-Null Count	Dtype
0	created_at	2331 non-null	object
1	id	2331 non-null	int64
2	id_str	2331 non-null	object
3	full_text	2331 non-null	object
4	truncated	2331 non-null	bool
5	display_text_range	2331 non-null	object
6	entities	2331 non-null	object
7	extended_entities	2059 non-null	object
8	source	2331 non-null	object
9	in_reply_to_status_id	77 non-null	float64
10	in_reply_to_status_id_str	77 non-null	object
11	in_reply_to_user_id	77 non-null	float64
12	in_reply_to_user_id_str	77 non-null	object
13	in_reply_to_screen_name	77 non-null	object
14	user	2331 non-null	object
15	geo	0 non-null	object
16	coordinates	0 non-null	object
17	place	1 non-null	object
18	contributors	0 non-null	object
19	is_quote_status	2331 non-null	bool
20	retweet_count	2331 non-null	int64
21	favorite_count	2331 non-null	int64
22	favorited	2331 non-null	bool
23	retweeted	2331 non-null	bool
24	possibly_sensitive	2197 non-null	object
25	possibly_sensitive_appealable	2197 non-null	object
26	lang	2331 non-null	object
27	retweeted_status	163 non-null	object
28	quoted_status_id	26 non-null	float64
29	quoted_status_id_str	26 non-null	object
30	quoted_status_permalink	26 non-null	object
31	quoted_status	24 non-null	object

3- **HTTP Request:** additional tsv formatted data can be accessed through HTTP request on the following URL : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv . data loaded from this source into **df_tsv** data frame with below structure :

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   tweet_id  2075 non-null   int64
 1   jpg_url   2075 non-null   object
 2   img_num   2075 non-null   int64
 3   p1        2075 non-null   object
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool
 6   p2        2075 non-null   object
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool
 9   p3        2075 non-null   object
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool
```

Assess

Data Assessment activity shows the following issues :

Data Quality Issues:

Validity:

- Tweet Data in df_CSV: 'timestamp' is formatted as string not datetime
- Tweet Data in df_JSON: 'created_at' is formatted as string not datetime

Accuracy:

- Tweet Data in df_CSV: rating_numerator below 15 is 2330 , 26 records found to have odd values like 420 , 72 ,....
- Tweet Data in df_CSV: rating_denominator should be 10 , but only 2333 records found to be 10 , 23 records found to have odd values like 0 , 170,.... found
- Tweet Data in df_CSV: data includes retweets and replys not only original tweets
- Tweet Data in df_CSV: doggo, floofer, pupper ,puppo variable have zero null values , however String 'None' is used instead .

Consistancy:

- Tweet Data in df_JSON : has 0 records with retweeted = True , however 2353 records found to have retweet_count > 0
- Tweet Data in df_JSON : has 8 records with favorited = True , however 2175 records found to have favorite_count > 0

Data Tidiness Issues:

- Tweet Data in df_CSV: Values of dog_stages 'doggo', 'floofer', 'pupper', 'puppo' are represented as Variable

- Image Data in df_TSV: p1, p2,p3 are three columns for the same variable for the gog breed
- Tweet Data in df_JSON: retweet is represented in df_csv.retweeted_status_id and df_json.retweeted

Clean

Validity:

- Tweet Data in** df_CSV**: format 'timestamp' as datetime

- Tweet Data in** df_JSON**: format 'created_at' as datetime¶

Accuracy:

-Tweet Data in df_CSV: rating_numerator and rating_denominator has outlier values

Solution : create a new Column 'rating' taht calculates the rating percentage (rating_numerator/rating_denominator) * 100 , so that high numerators with high denomerators will yeild a normal percentage , oullier still there but this will be handled using outliers techniques in reporting .

- Tweet Data in df_CSV: data includes retweets and replys

solution : delete records 'having retweeted_status_id' or 'in_reply_to_status_id' not null

- Tweet Data in df_CSV: doggo, floofer, pupper ,puppo variable have zero null values , however String 'None' is used instead .

solution : Clean doggo, floofer, pupper ,puppo variable (the last four columns) by replacing 'None' with ''

- Image Data in** df_TSV**: out of 2075 , 543 record found to have 'isdog' is false with undesired values fo the 'breed' column like :'pug', 'beaver','envelope','bakery',....

Solution:

- consider only the record with isdog = true , and set the remaining to default value (i.e. 'Unknown')

Consistancy:

- Tweet Data in df_JSON : has 0 records with retweeted = True , however 2353 records found to have retweet_count > 0

- Tweet Data in df_JSON : has 8 records with favorited = True , however 2175 records found to have favorite_count > 0

Solution : remove retweeted variable and favorited column and use retweet_count and favorite_count to judge if the tweet is favirited or retweeted

Data Tideness:

- Tweet Data in df_CSV: Values of dog_stages 'doggo', 'floofer', 'pupper', 'puppo' are represented as Variable

Solution :

- Add dog_type by concatenation 'doggo', 'floofer', 'pupper', 'puppo'

- recurseviely clean the '-'

- Finally drop the columns¶

- Tweet Data in *df_TSV*: ***p1, p2,p3 are three columns for the same variable for the breed***

Solution:

- consider only the record with Max 'conf' value for each tweet and ignore the others¶

- Tweet Data as a single ovservational unit is not is a single table : Tweet Data is spreaded over df_JSON , df_TSV and df_CSV .

Solution :

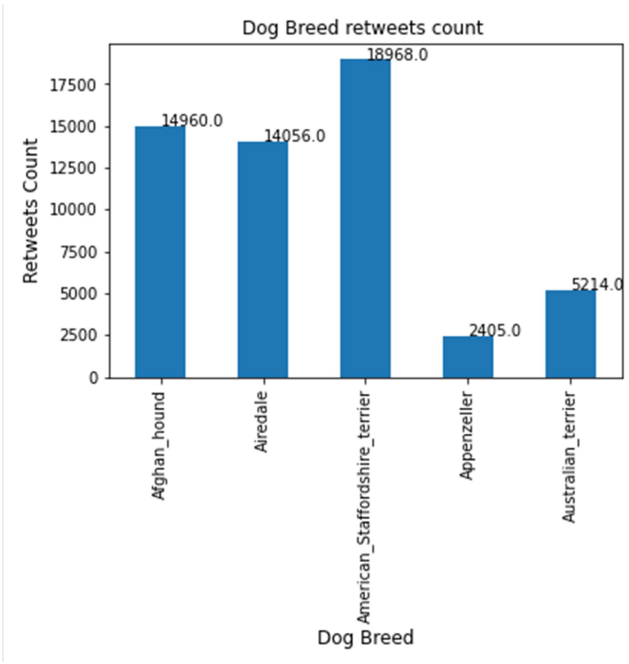
merge 'retweet_count','favorite_count' , 'dog_breed' data into df_CSV using tweet_id¶

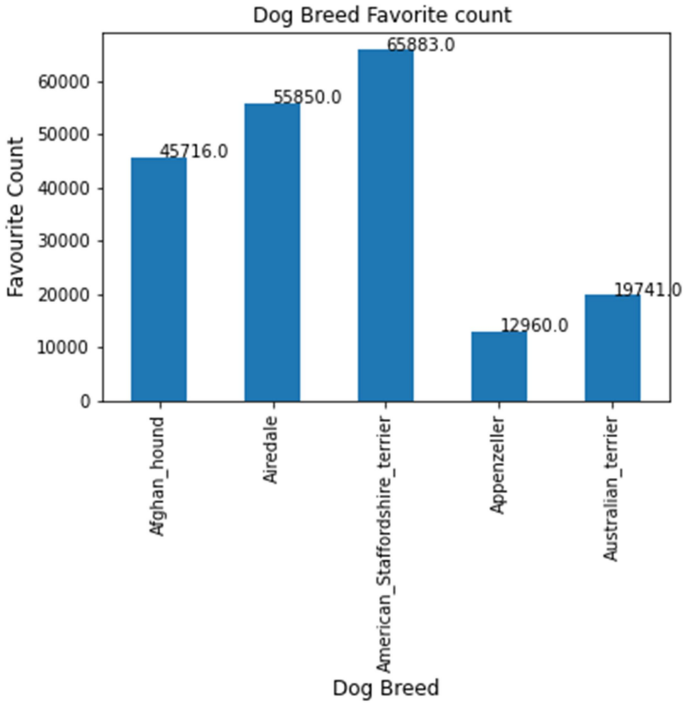
Master data frame for Tweets and image :

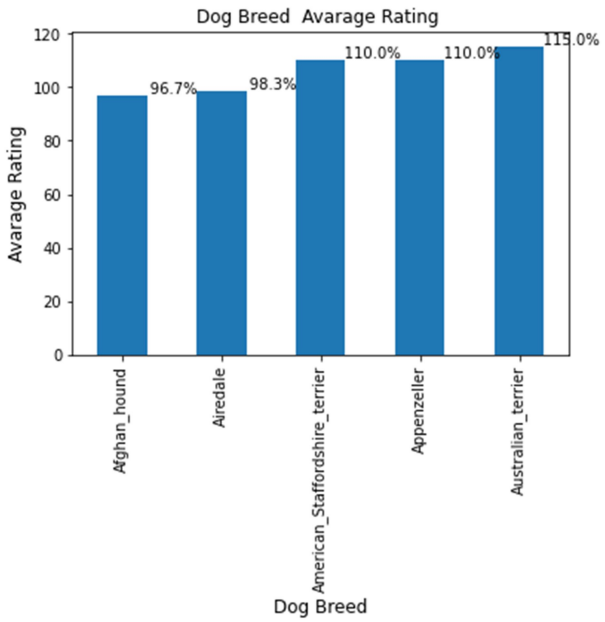
- Finally we store a master data frame for the cleaned data of Tweets and Images in to file named : "twitter_archive_master.csv"

Visualization : Building Diagrams for data insights :

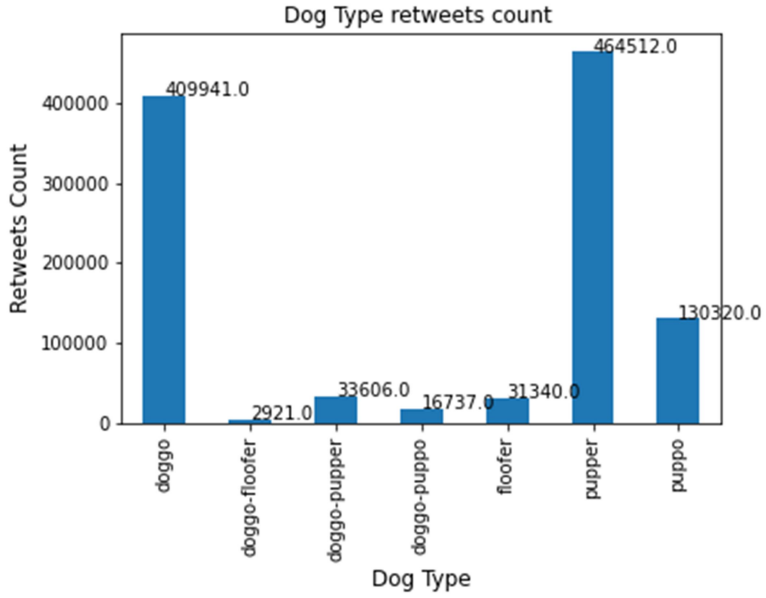
- ***Top 10*** *Dog Breeds* ***retweeted and top 10 breeds marked as Favorite***

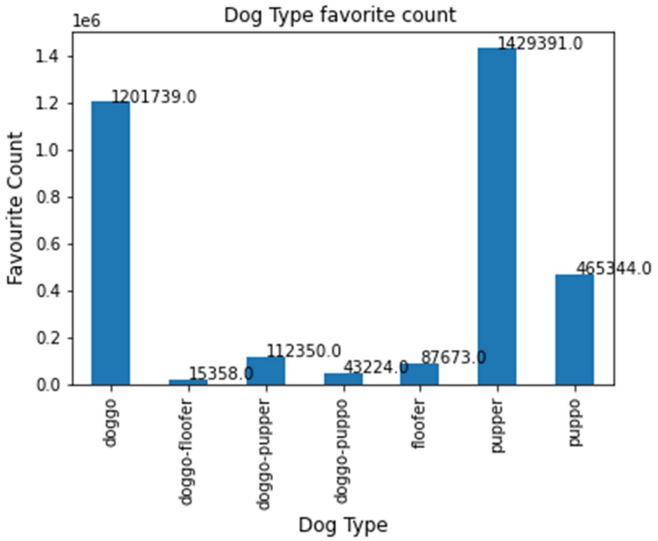

Dog Breed retweets count
Retweets Count
0
2500
5000
7500
10000
12500
15000
17500
14960.0
14056.0
18968.0
2405.0
5214.0
Afghan_hound
Airedale
American_Staffordshire_terrier
Appenzeller
Australian_terrier
Dog Breed

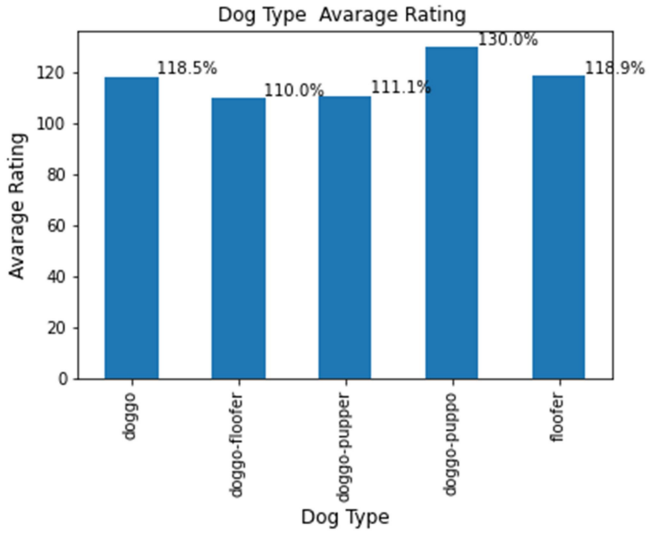

Dog Breed Favorite count
Favourite Count
0
10000
20000
30000
40000
50000
60000
45716.0
55850.0
65883.0
12960.0
19741.0
Afghan_hound
Airedale
American_Staffordshire_terrier
Appenzeller
Australian_terrier
Dog Breed



- ***Top 10*** *Dog Types* ***having retweets and top 10 marked as Favorite¶***







Conclusion:

from the above analysis we found the the following Facts

the top rated dogs are :

- Pupper

- Doggo

- Puppo

the top rated dog breeds are :

- American_Staffordshire_terrier

- Afghan_hound & Airedale

- Australian_terrier¶