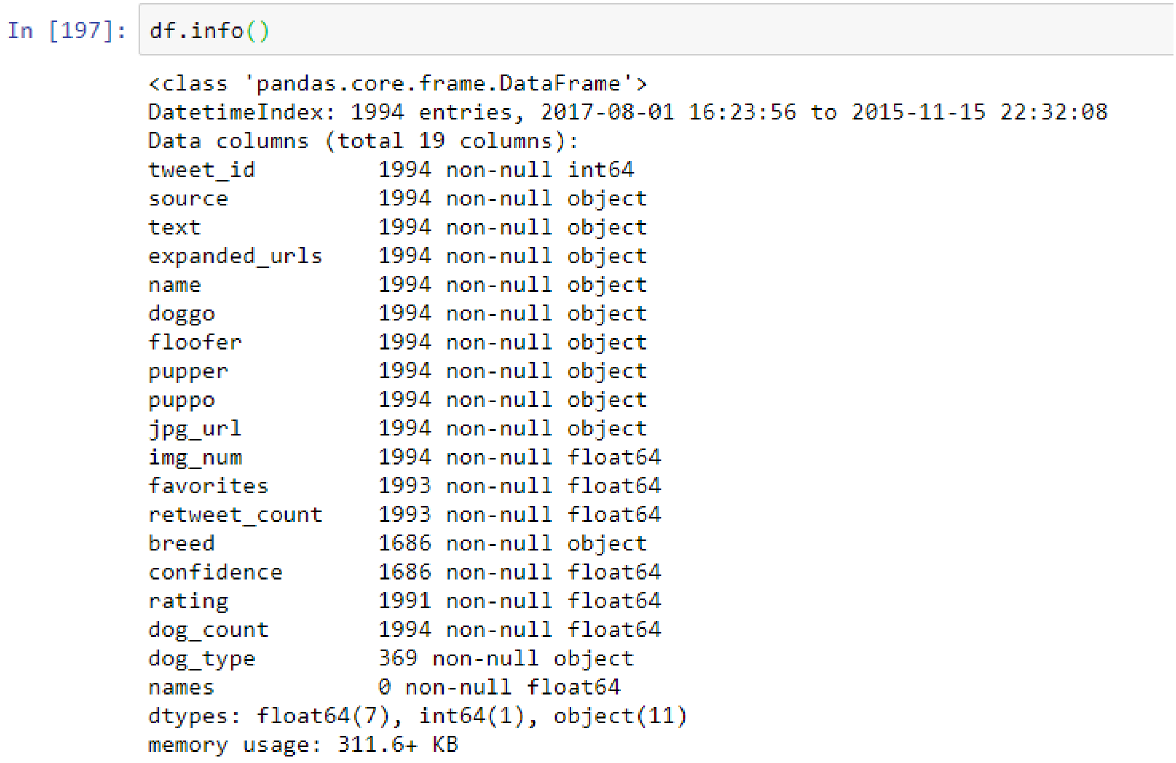
Data Story – by Thanh Dinh

There is a huge effort which has been done to bring up the cleaned, meaningful, final dataset. The next step could be state of art, which tell a story about dog rating, based on twitter_archive_master.csv

Everyone will have a dog to love. But, which dog type (breed) is most common dog in the tweet dataset? What dog type has the highest average rating? These questions and more are answered in the following insights.

1. The final data

There are some qurery about the final data:



```
In [197]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1994 entries, 2017-08-01 16:23:56 to 2015-11-15 22:32:08
Data columns (total 19 columns):
tweet_id          1994 non-null int64
source            1994 non-null object
text              1994 non-null object
expanded_urls     1994 non-null object
name              1994 non-null object
doggo             1994 non-null object
floofer           1994 non-null object
pupper            1994 non-null object
puppo             1994 non-null object
jpg_url           1994 non-null object
img_num           1994 non-null float64
favorites         1993 non-null float64
retweet_count     1993 non-null float64
breed             1686 non-null object
confidence        1686 non-null float64
rating            1991 non-null float64
dog_count         1994 non-null float64
dog_type          369 non-null object
names             0 non-null float64
dtypes: float64(7), int64(1), object(11)
memory usage: 311.6+ KB
```

2. The correlation of Retweets and Favourites Fig1

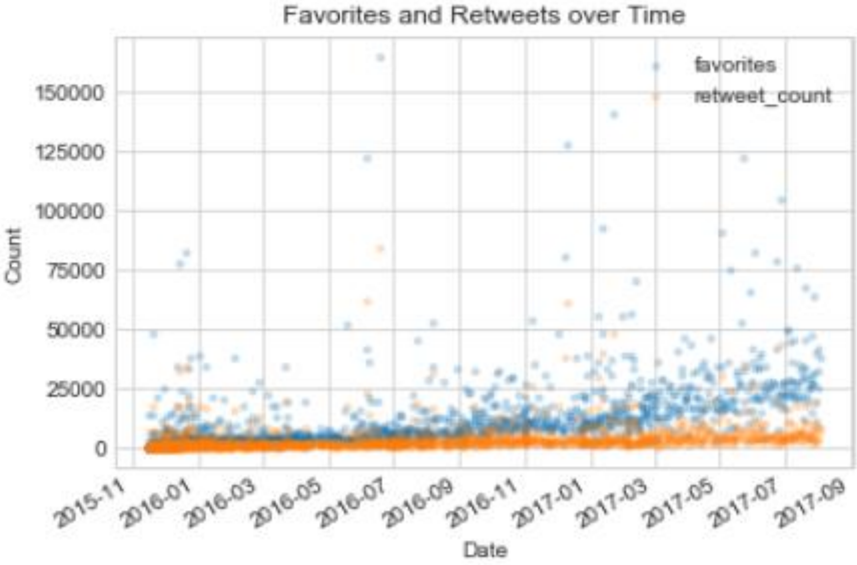


Fig1: Relation of re-tweet & favourite

Fig 1 represent number of favourite (blue) compare with re-tweet (orange) over the time. The number of re-tweet & Favourites are incline time after time. From my point of view, the trend of data shows the popularity of tweeter accounts. And, breed topic is hot for quite a lot of people over internet

3. Histogram of Most the Rated Breeds

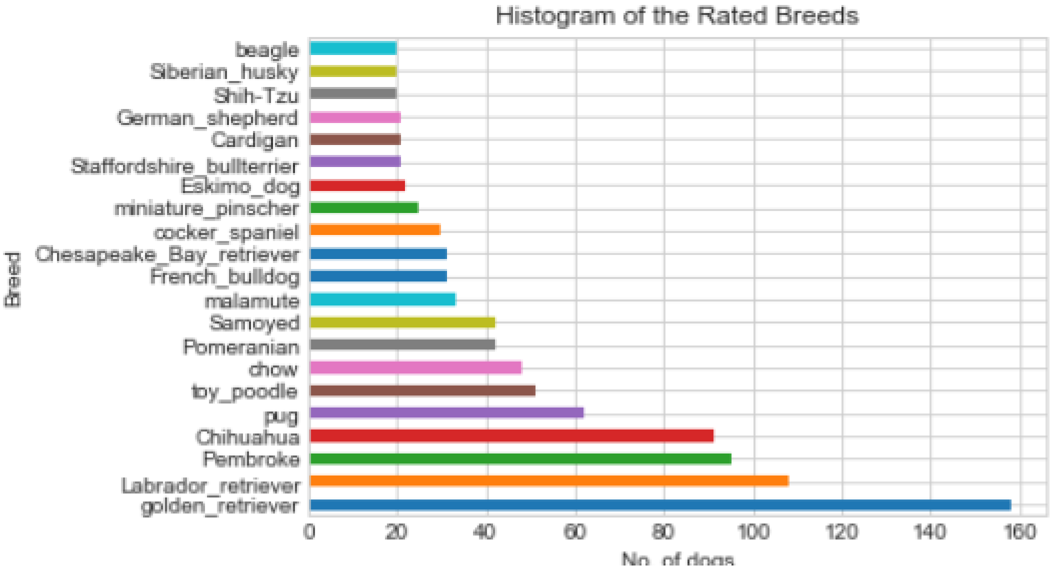


Fig2: Rate Breeds histogram

Fig 2 shows the ranking of breeds which counts from final dataset. Un-surprising, golde_retriever is the most wanted breeds with over 150 rates, far lag behind the runner (Labrador_retriever). Sadly, my dog (Shih-Tzu) deserve the lowest points (around 20). Anyhow, just because I do not use Tweet much.

4. Outlier consideration

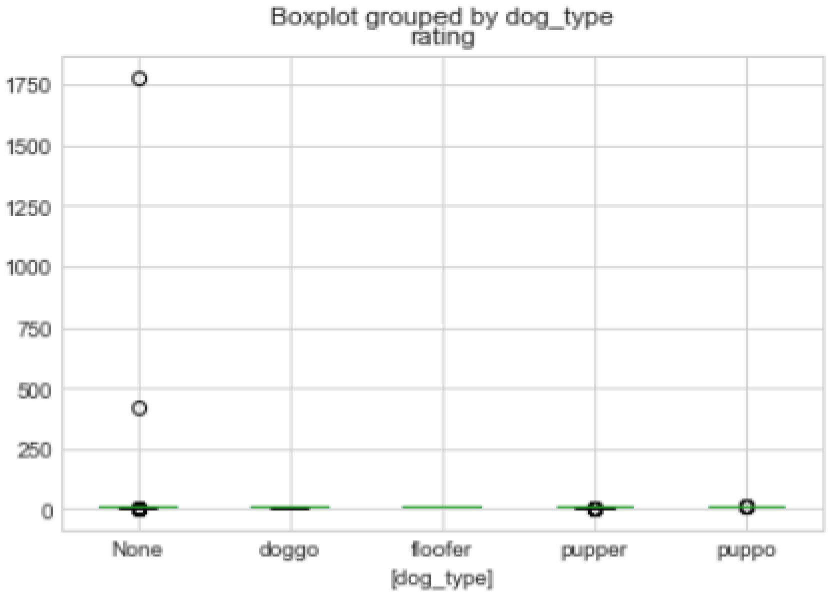


Fig3: Boxplot of dog type vs rating. (overrate)

It must be something typo with the rate number or someone overrating their dog. To fix this issue, I get rid of those rates over 14.

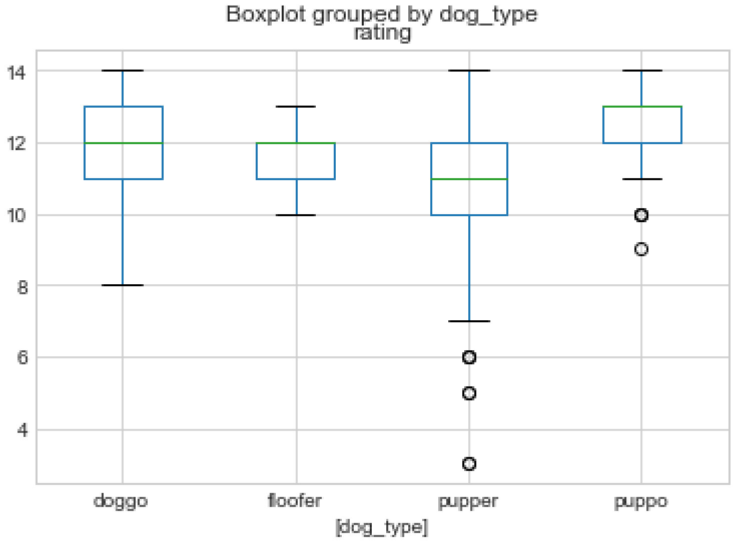


Fig4: Boxplot of dog type vs rating