Wrangling Act Report

# For external use only

This report is intended to document analysis process and insights found afterwards.

The original dataset contains a lot of quality and tidiness issues, after careful wrangling process was performed on it, the new cleaned dataset is presented as following and analysis process starts from examining dataset info with *.info*() method. The resulted details are shown below

<class 'pandas.core.frame.DataFrame'>

Int64Index: 1746 entries, 0 to 1745

Data columns (total 13 columns):

breed 1746 non-null object

tweet\_id 1746 non-null object

timestamp 1746 non-null int64

source 1746 non-null object

text 1746 non-null object

name 1746 non-null object

doggo 1746 non-null object

floofer 1746 non-null object

pupper 1746 non-null object

puppo 1746 non-null object

favorite\_count 1746 non-null int64

retweet\_count 1746 non-null int64

rating 1746 non-null float64

dtypes: float64(1), int64(3), object(9)

memory usage: 191.0+ KB

Table 1.1 dataset info details

A few examples of instances from dataset is shown below

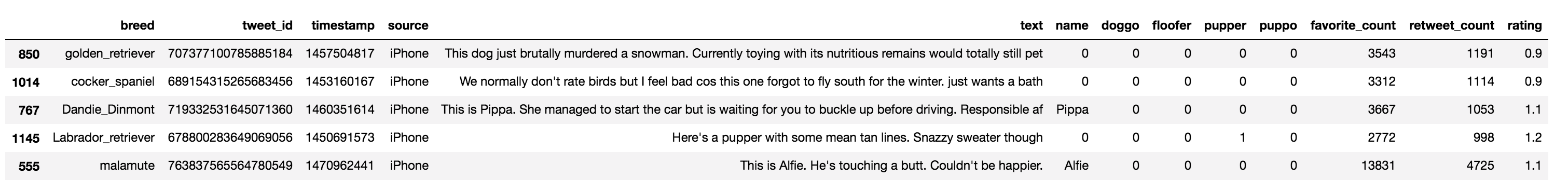


Figure 1.1 a few samples from dataset

After further analysis, that three insights are revealed:

1. The majority users tweet through iPhone app
2. Most tweeted pictures are of golden retriever
3. Most popular beloved tweets are about Pomeranian
4. By first sight, source column contains simple data of where the user tweeted from, further examining entire column shows that the majority of users used twitter on iPhone (Figure 1.2).



Figure 1.2 Chart of source column

1. breed column contains predicted dog breeds of all tweets, it suggests what kind of dogs in pictures people tweeted. The top 10 breeds as figure 1.3 indicates, Golden retriever appears to be in most popular, Labrador retriever takes second place. As a matter of fact, retrievers are still most common human friend among all dog breeds.

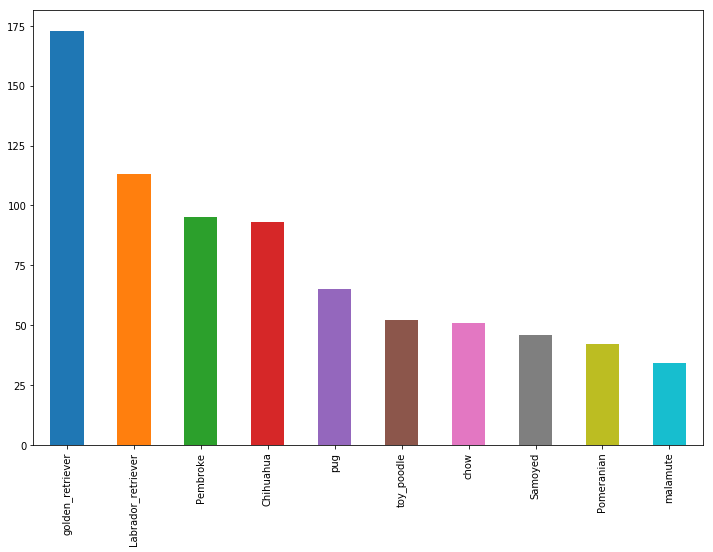


Figure 1.3 Top 10 dog breeds

1. Furthermore, retweet count and favorite count both indicate that how much people love certain tweets. By comparisons it uncovers Pomeranian to be the most beloved dog among twitters.

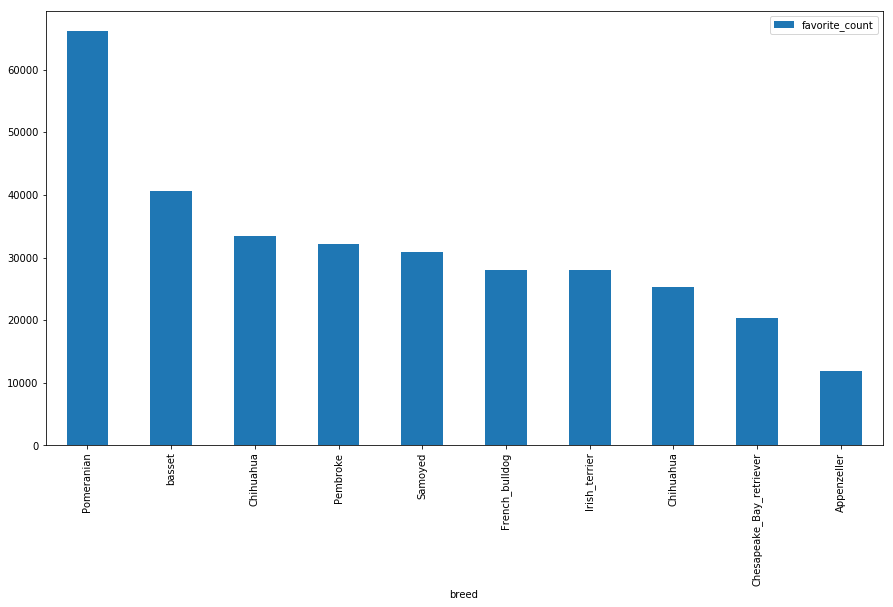


Figure 1.4 favorite\_count by breed

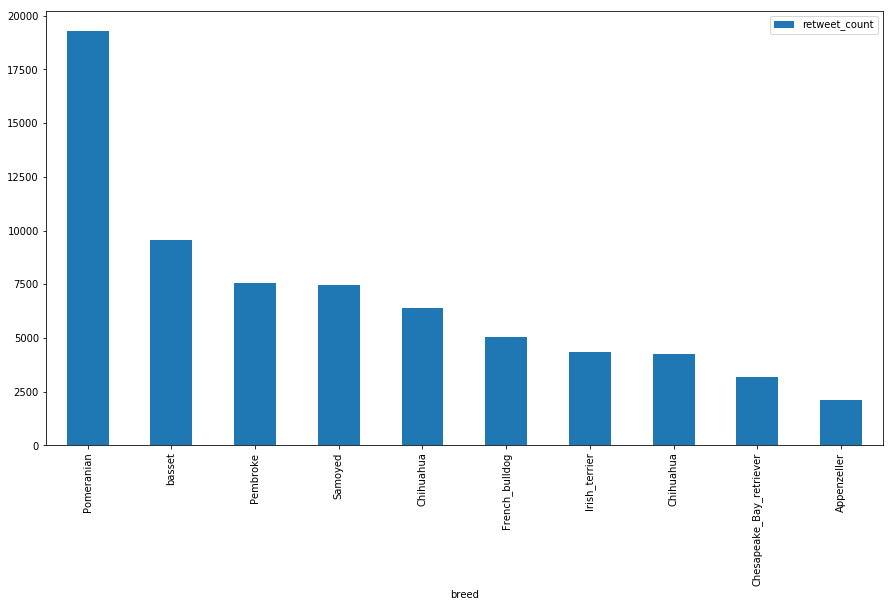


Figure 1.5 retweet\_count by breed

In a word, meaningful insights can be derived from analyzing cleaned data, but original dataset with noises has to be wrangled and it is the most essential element in order to complete further data analysis.