Wrangling @WeRateDogs Twitter Archive

By: Divya Prasanth P

What is Data Wrangling?

Data Wrangling is the process of gathering, assessing, cleaning and storing data. These steps are intended to make the raw_data functionally fit for data analysis to find hidden insights. These insights in the data help us make data-oriented decisions.

Let us go through these data wrangling steps one by one and find insights in the data.

1. Data Gathering:

   Raw data is gathered in this step. Data gathering can be done through various methods. These methods include:
   1. Downloading data directly via any source
   2. Downloading data programmatically using python packages. For this project I used requests library
   3. Accessing data of websites using Application Programming Interfaces(APIs). For this project, using tweepy library I accessed @WeRateDogs twitter archive

2. Data Assessing:

Assessing data can be done in two ways: Visually and Programmatically

1. Visual Assessment:

Visual assessment is usually done to acquaint oneself with the data. We see the data in its entirety and understand the data. Understanding each column in the data is very important for further assessment.

2. Programmatic Assessment:

Using programmatic tools such as pandas,descriptive statistics, matplotlib to assess the data. This type of using code to assess the raw data is known as programmatic assessment. Using programmatic assessment we could try to find the following observations.

1. If there is any missing data in the data set
2. Data type of each variable in the data set
3. Checking for duplicates
4. Checking for invalid entries
5. Checking for data consistency, etc

3. Data Cleaning:

In this step, the observations made in the data assessing step are transformed into action items. These action items are then implemented via code to clean the raw_data. Data cleaning involves converting observations made in the data assessment into pseudocode(define), then coding(code) to clean the code, and then testing to check if the particular observation is resolved. At the end of this step, after all the observations have been resolved using define>>code>>clean, the resulting data is an enriched data functionally fit for further data analysis.

4. Data Storing:

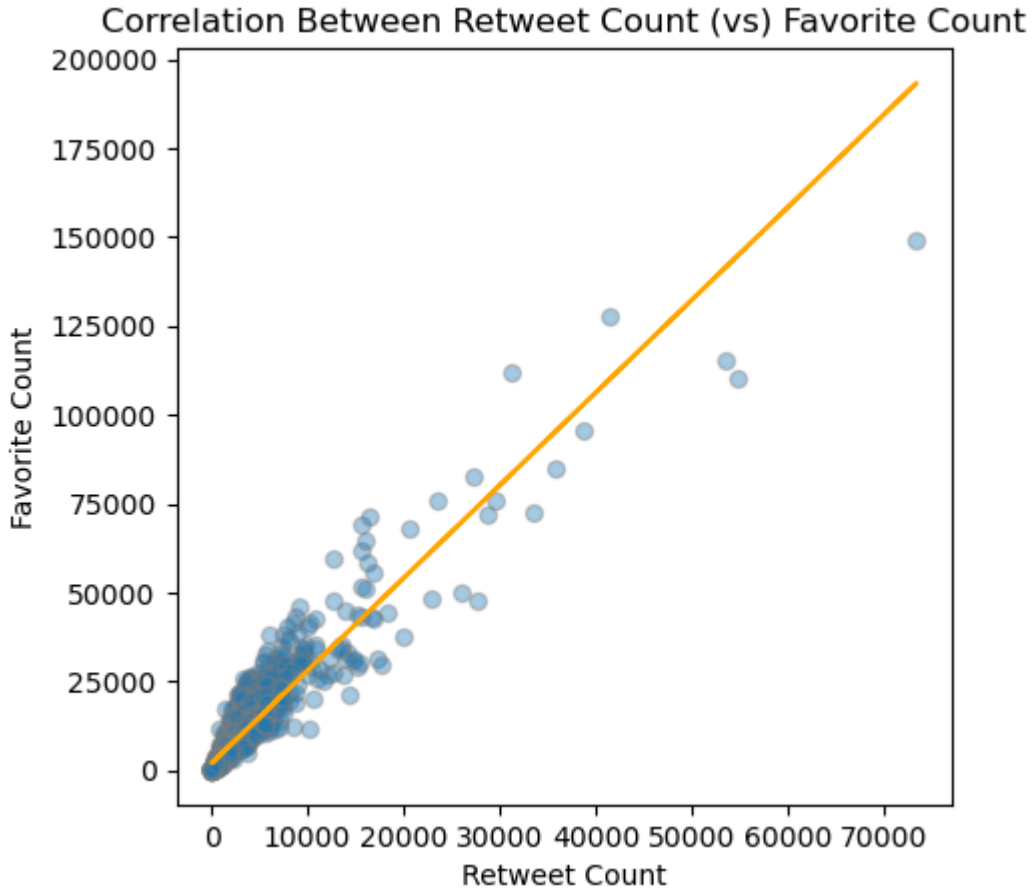
This step involves saving the cleaned data as a file, or to a database.

I took @WeRateDogs twitter account archive and tried to find some insights. Initially I had like approximately 2500 tweets which contained only basic information of the tweet_id, dog_name, dog_stage, rating_numerator, rating_denominator. I followed the above data wrangling process and came up with a high quality and neat data. The resultant dataset contained 1946 tweets with additional data such as retweet_count, favorite_count, and predicted dog_breed.

The following are some interesting insights that I found through EDA(Exploratory Data Analysis) in the datasets.

Insight 1

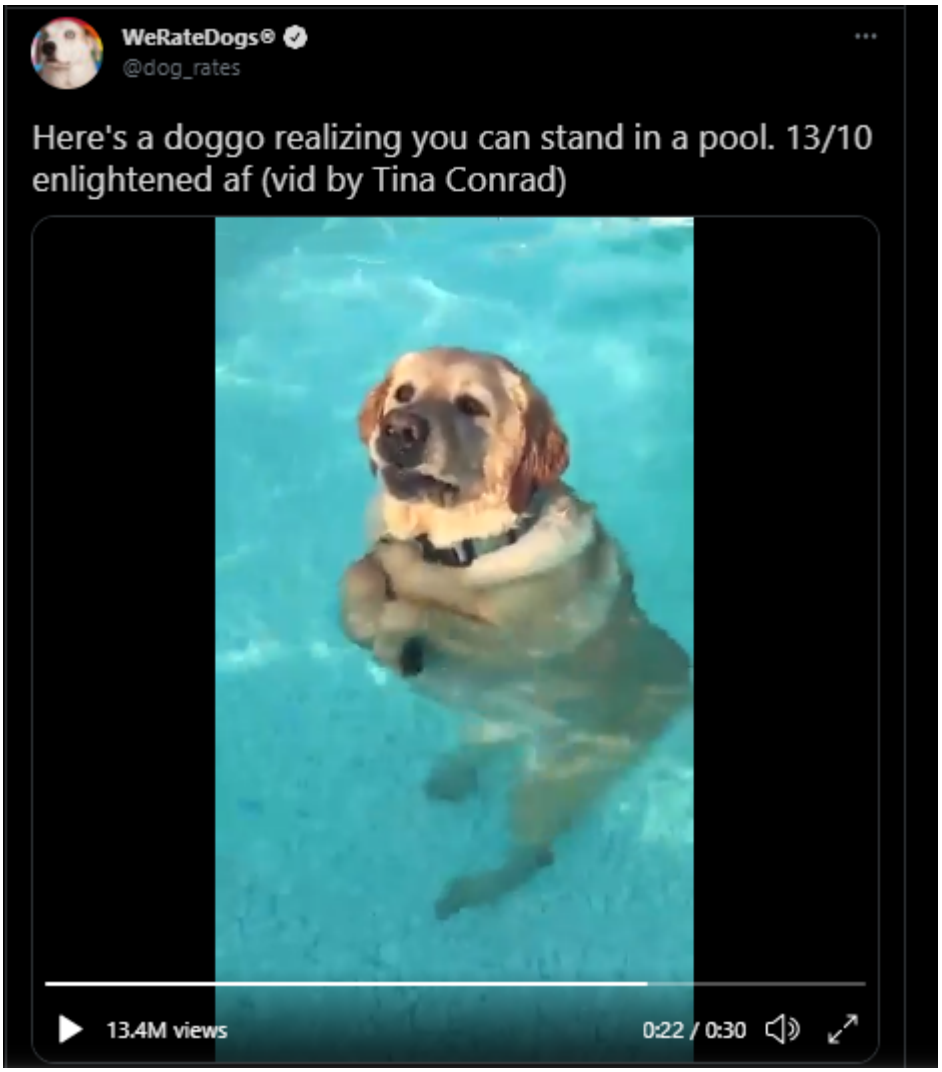
Is there any correlation between retweet count and favorite count?



1. There existed a strong positive correlation between ‘Retweet Count’ and ‘Favorite Count’.
2. As ‘Retweeted Count’ increases, ‘Favorite Count’ also increases and vis-a-vis
3. To note: these both variables have high correlation but not causation. Ie, increase in one does not cause an increase in the other.

Insight 2

Which tweet had the most likes?



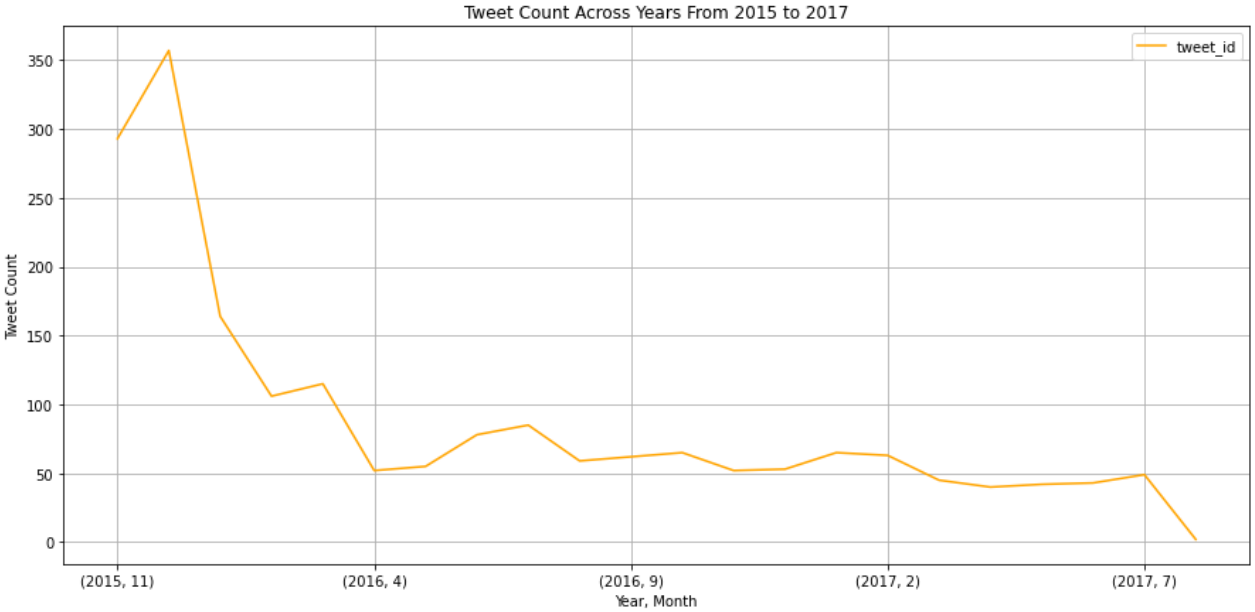
Yeah!!!! I know! The above dog is cute <3. This tweet had:

1. The above tweet was retweeted 73.4K times(to date)
2. It has got above 149K likes.

The above insight got me interested and I tried to find the dog_breed which has the most likes. Before that let's see how many tweets were posted by the @WeRateDogs through the years.

Insight 3

How has the tweets at @WeRateDogs account been over the years?

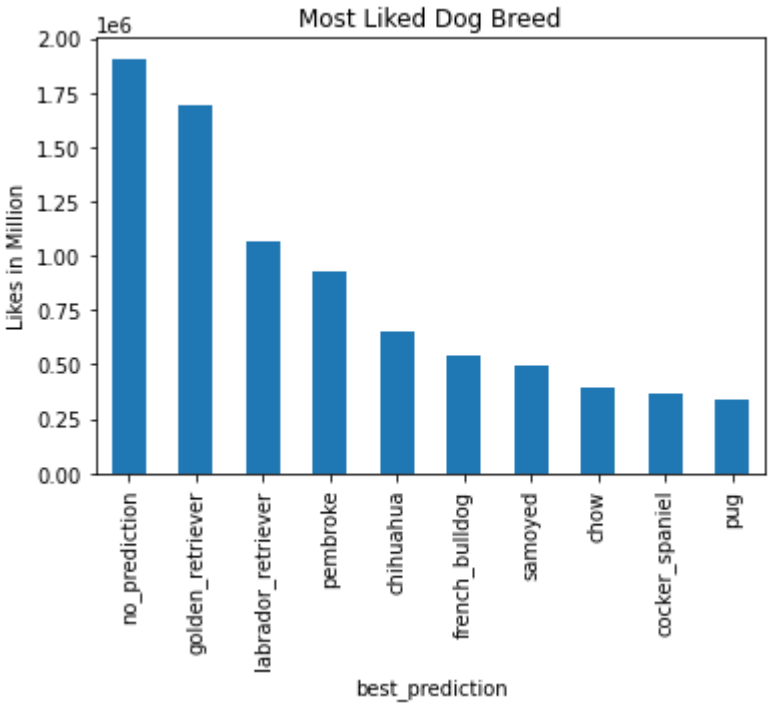


From the above line graph, in the available data, we could see that, @WeRateDogs twitter account:

1. has posted the most number of dog_ratings around the 2015 Year end and early 2016
2. After mid 2016, there has been a decrease in the tweets

Insight 4

Which breed has been most popular?



From the above bar graph, we could see that:

1. "no_predictions" indicate tweets that predicting models couldn't predict as dogs
2. Apart from "no_predictions" golden_retriever is the most liked dog breed
3. labrador_retriever stands next and so on