WeRateDogs Tweet Data Wrangling Project

For the purpose of this project, tweet archived data of Twitter user @dog_rates, also known as WeRateDogs was utilised. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. They have over 4 million followers. The project explored their basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

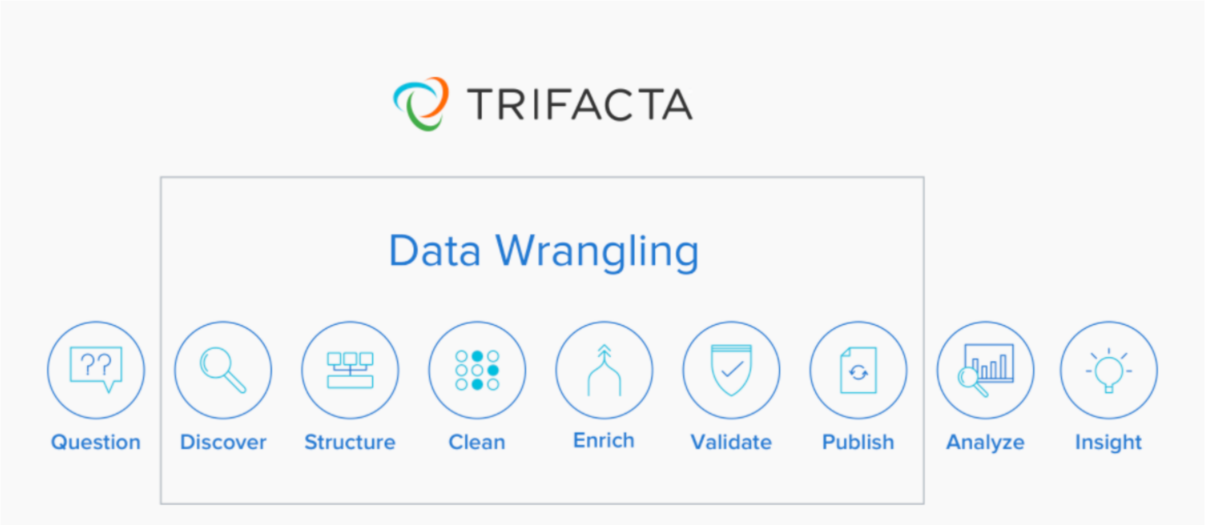
WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Image via ***Boston Magazine***

With the two dataframes that were created, one that was a concat of the twitter-archive-enhanced.csv and tweet-json.txt files while the second dataframe was from the image_predictions.tsv file, I only used the first dataframe for the exploratory analysis. Following were the visuals created and a brief description:

What is Data Wrangling?

Data wrangling is the process of gathering, selecting and transforming data to answer an analytical question. It is also known as data cleaning and also the step where 80% of the time is spent by analytics professional.(ref: www.elderresearch.com)



For the purpose of this project:

- Gathering data
- Assessing data
- Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) my data wrangling efforts and 2) my data analyses and visualizations

Gathering Data

There were three data sources that was utilized:

1) First was the "Twitter Archive" data that had 5000+ of @WeRateDogs tweets that was provided already

2) Second was the data pull by using Twitter's API. Any other data that was not available in the "Twitter Archive data" was extracted from the txt file in json format.

3) Third was an "Image Predictions" file that contained information on type of breed the dog is. This was generated through a machine learning model that classified the breed of dogs.

Assessing Data

The goal of this steps is to identify any quality and tidiness issues. Assessment of data is done in two steps:

1) First is the visual assessment where you eyeball the tables.

2) Second is the programmatic assessment where you use several functions such as info(), describer(), shape and so on.

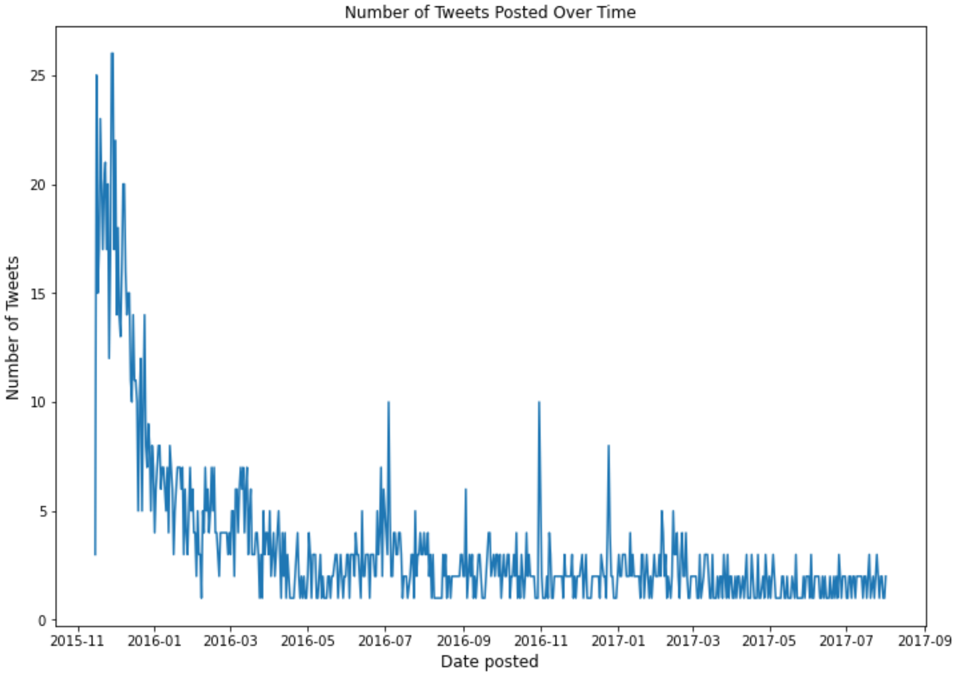
A good principle is to document the issues identified as you go about assessing the data.

Cleaning Data

Cleaning data involves programmatically cleaning the issues that were identified while assessing the data.

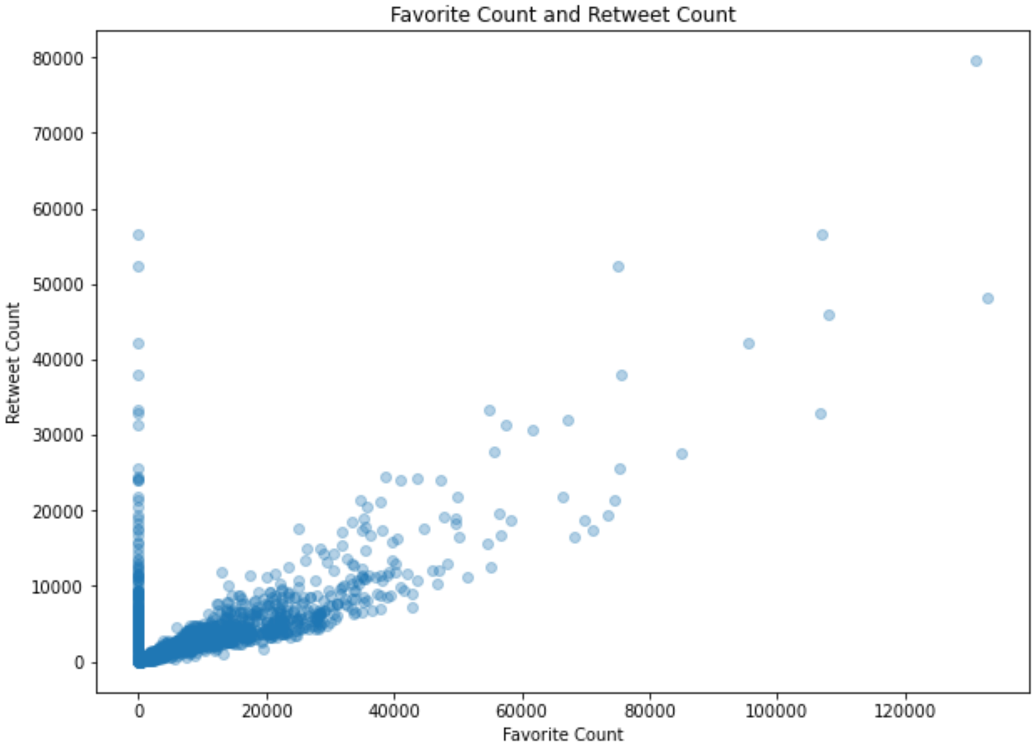
Documentation and Analysis of Final Data

Visual 1: Number of Tweets Posted Over Time



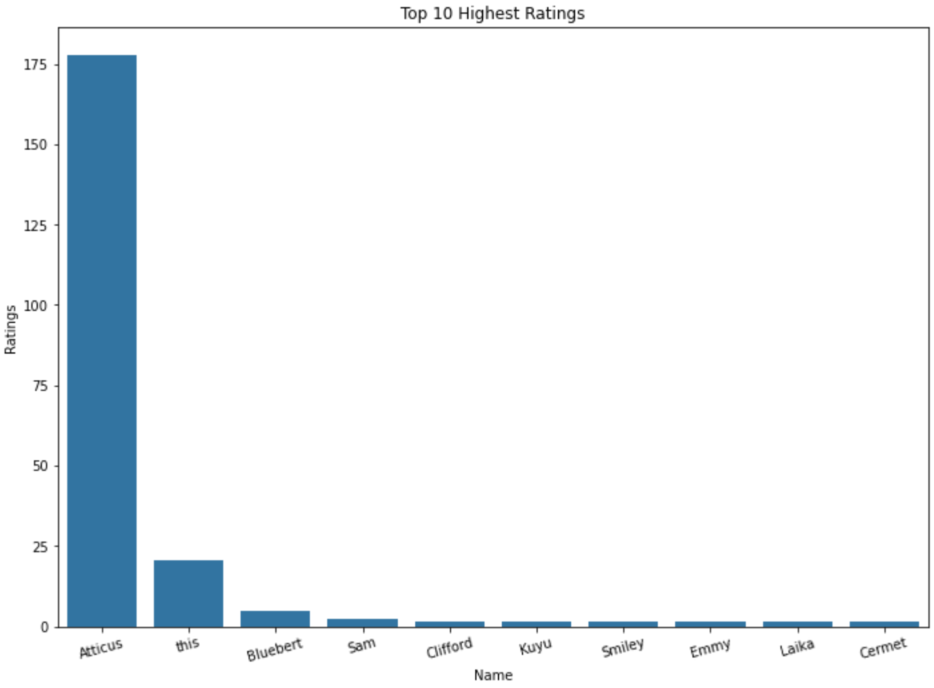
Description: The line chart shows the activity of the @WeRateDogs over time. As seen from the above line chart, there has been decrease in the no. of postings over time. The account was more active posting tweets during late 2015 and early 2016. The highest number of tweets posted by the account is approximately 27.

Visual 2: Favorite Count and Retweet Count



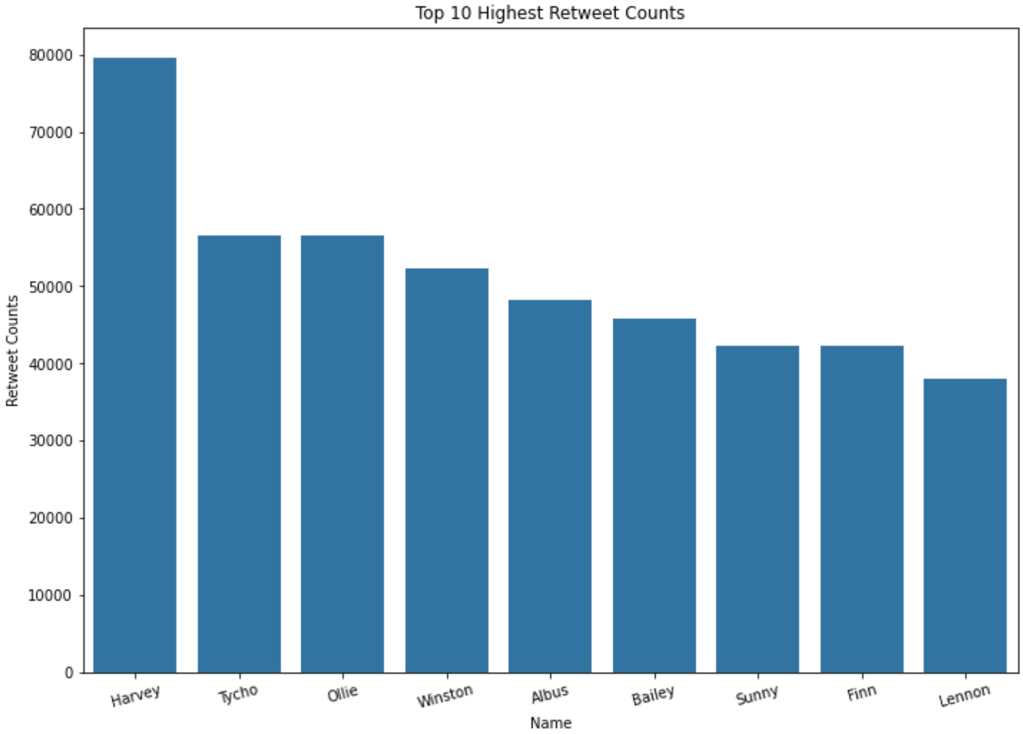
Description: The scatter plot has been constructed with the favorite count along the x-axis and retweet count along the y-axis. Few outliers were observed which can distort our analysis. For the purpose of our analysis, these outliers will not be removed as we would like to visualise those with the highest receiver of ratings and count.

Visual 3: Top 10 Highest Rating



Description: The bar chart has been plotted against ratings which is the results of rating_numerator divided by rating_denominator. Only top 10 ratings were visualised since there are many unique names and visualising all the names could crowd the bar chart. Therefore, only top 10 ratings were displayed. From the bar chart, it can seen that the highest rating receiver goes by the name of Atticus who received a rating of 175.

Visual 4: Top 10 Highest Retweet Counts



Description: I wanted to check if there were certain stages of dogs that were popular among the audiences. From the above bar chart, it can be seen that the most popular retweeted dogs were at the stage "puppo" and "doggo,floofer".

All in all, I think there were other features that could be visualised too but I only added 4 to keep it short and sweet and find out some of the interesting stuffs that I was curious about. So far we know that the account has been maintaining a steady postings over the months following 2016. Atticus received the highest ratings however, Harvey was more popular among the audiences. Last but not least, the stage of dogs that were popular among audiences were puppo and doggo,floofer.