We Rate Dogs – Udacity's Data Wrangling Project

Introduction:

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs assigns a score out of 10 for the dog (like showed in Figure 1 below). The scope of this project is to gather the data, clean the data and analyze the data. The data is gathered and cleaned using standard Python libraries.

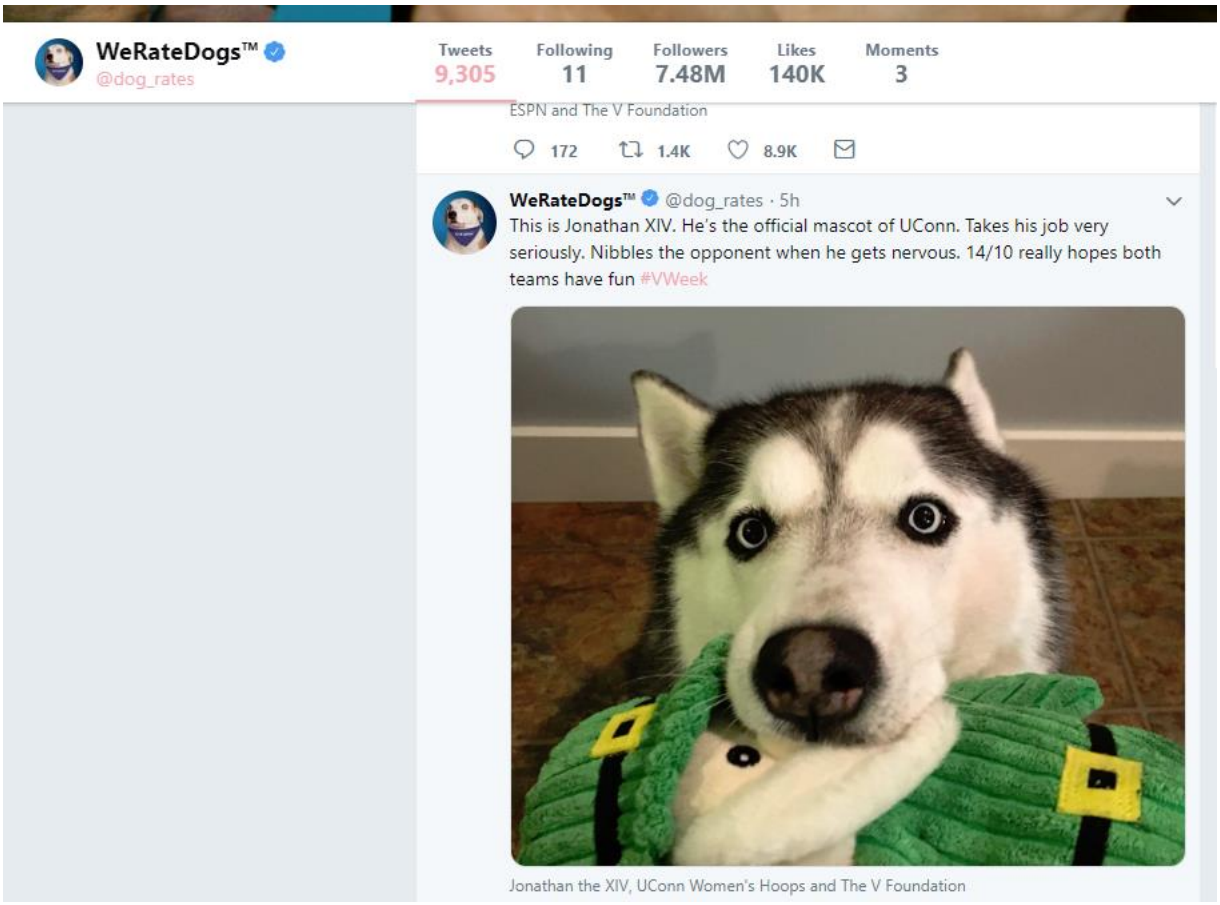


Figure 1: We Rate Dogs Tweet

Data Sources:

1. Twitter Archive (CSV) – This is a file that was sent to Udacity via email to be used for this project. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
2. Data from Twitter API – Data gathered from Twitter's API to obtain retweet count and favorite count information for the tweets.
3. Image Predictions File – Data to be downloaded from a URL which contains the image URL and prediction scores of a neural network.

Data Gathering:

1. The twitter Archive file was downloaded from Udacity's website and imported using Pandas' 'read_csv' function and stored in a data frame

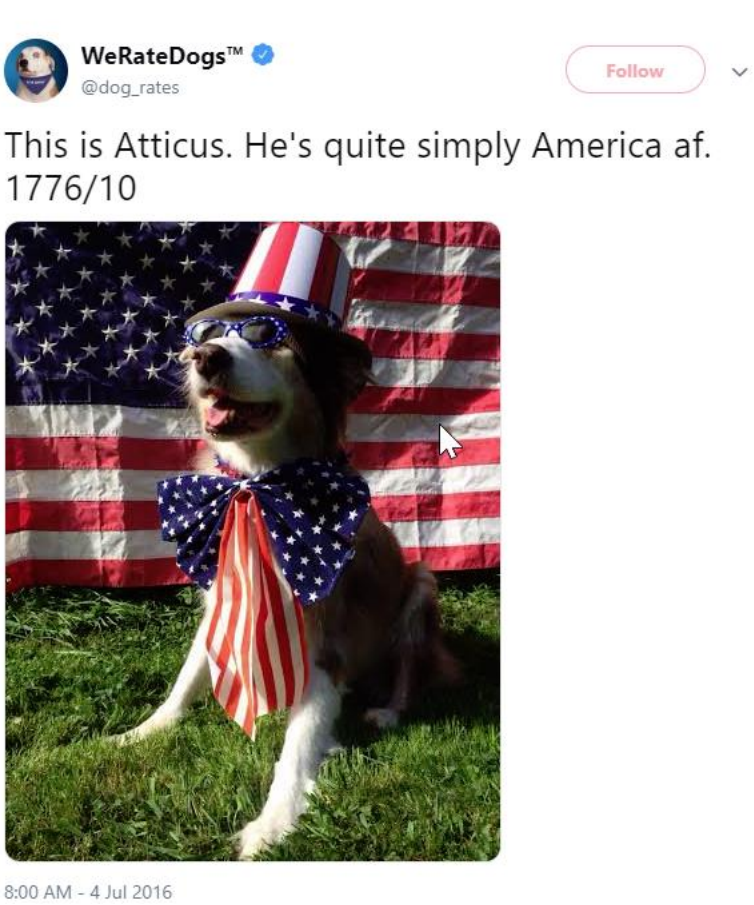
2. To obtain twitter data, the first step was to set up a developer account. Next, upon approval, got the Access Token and Access Token Secret keys which provided access the data. The data was in JSON format and each line (tweet) was stored in a file. The contents were then appended to a list. 'Tweet_Id', 'Retweet count' and 'Favorite count' were extracted and stored into a data frame.
3. The files were read programmatically from the URL using 'requests' library and stored in a data frame.

Data Assessment:

Upon assessing the data, there were quality and tidiness issues identified. These issues had to be fixed in order to proceed to analysis.

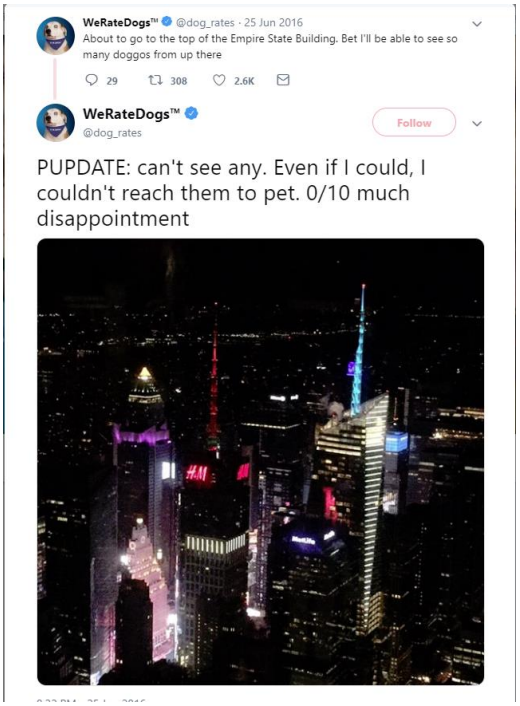
Quality Issues:

- There were retweets which had to be removed.
- The denominator had values other than 10, which had to be removed.
- The data types of a few columns had to be changed.
- There were a few tweets with scores like '1776' and '420' which had to be removed.



- The score was not captured correctly and decimals were ignored, which needed to be fixed.
- There were tweets classified into more than 1 dog stage, those were removed for higher quality analysis.

- There were tweets which had a 0 score as the tweet had an image of a street light instead of a dog, which had to be removed.



- The characters in the column was truncated, that had to be fixed.

Tidiness Issues:

- Dog stages are its own column. Needs to be condensed into 1 column.
- The data required for analysis can be merged into a single dataset using tweet_id as key column, instead of 3 data sets.

Data Cleaning:

There were copies created of the data frame to make sure that the original data was not affected. Standard Python libraries were used to clean the data. After every step of cleaning, there was testing done to ensure no unintended change had occurred. After the issues listed in the analysis section were resolved, the data was exported to a CSV file for further analysis using Tableau.

Author: Ajith Sharma