Data wrangling and analyzing project

Wrangle and Analyze Data

By: Mohamed fathy

Act Report

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

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10,

12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.
WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Project Goal

wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is done for "*Wow!*"-worthy analyses and visualizations.

Project Details

We worked on this project as follows:

- Data wrangling, which consists of:
    - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

Gather:

This project gathered data from the following sources:

The WeRateDogs Twitter archive. The ‘twitter-archive-enhanced.csv’ file was provided to Udacity Students .

- This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets.

- The Tweet image prediction, i.e., what breed dogs (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students (Like me).

- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favourite count at minimum and any additional data I find interesting.

Assessing Data

Once the data was gathered, I began to assess the data on both quality and tidiness issue.
There are four main issue in quality dimensions:

1. Completeness: Missing data
2. Validity: Does the data make sense
3. Accuracy: Inaccurate data
4. Consistency: Standardization

And There are three main requirements for tidiness:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observation unit forms a table

Clean Data

Cleaning data is tedious and often iterative. Just when data analyst believe they found all quality and tidiness issue, they often found additional issue arises. The cleaning process involves three steps.

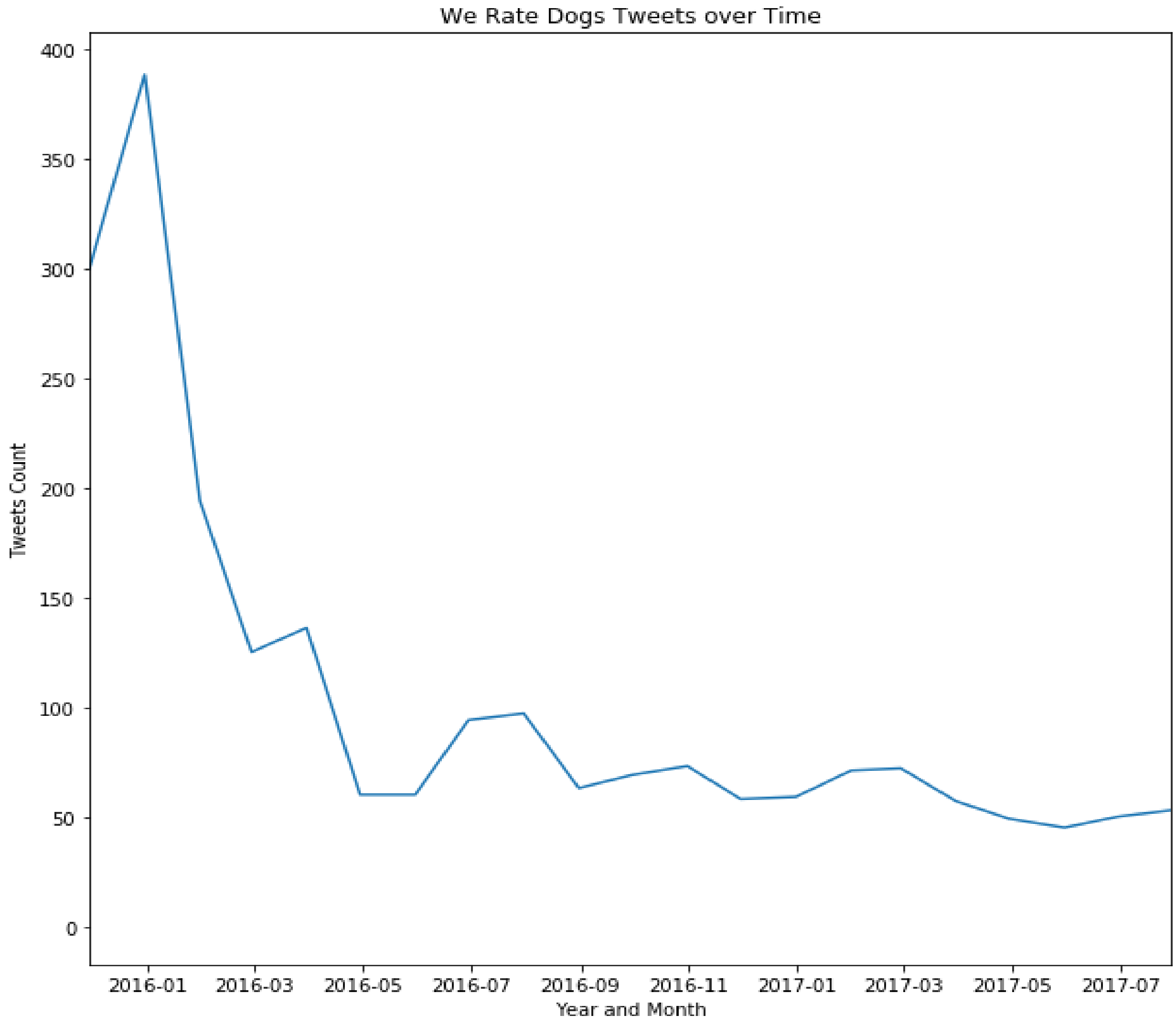
1. Define: Determine exactly what needs to be clean and how.
2. Code: Programmatically clean the code
3. Test: Evaluate the code to ensure the data set was cleaned properly.

Analysis and Visualization

There are several analysis and questions we answered as following :

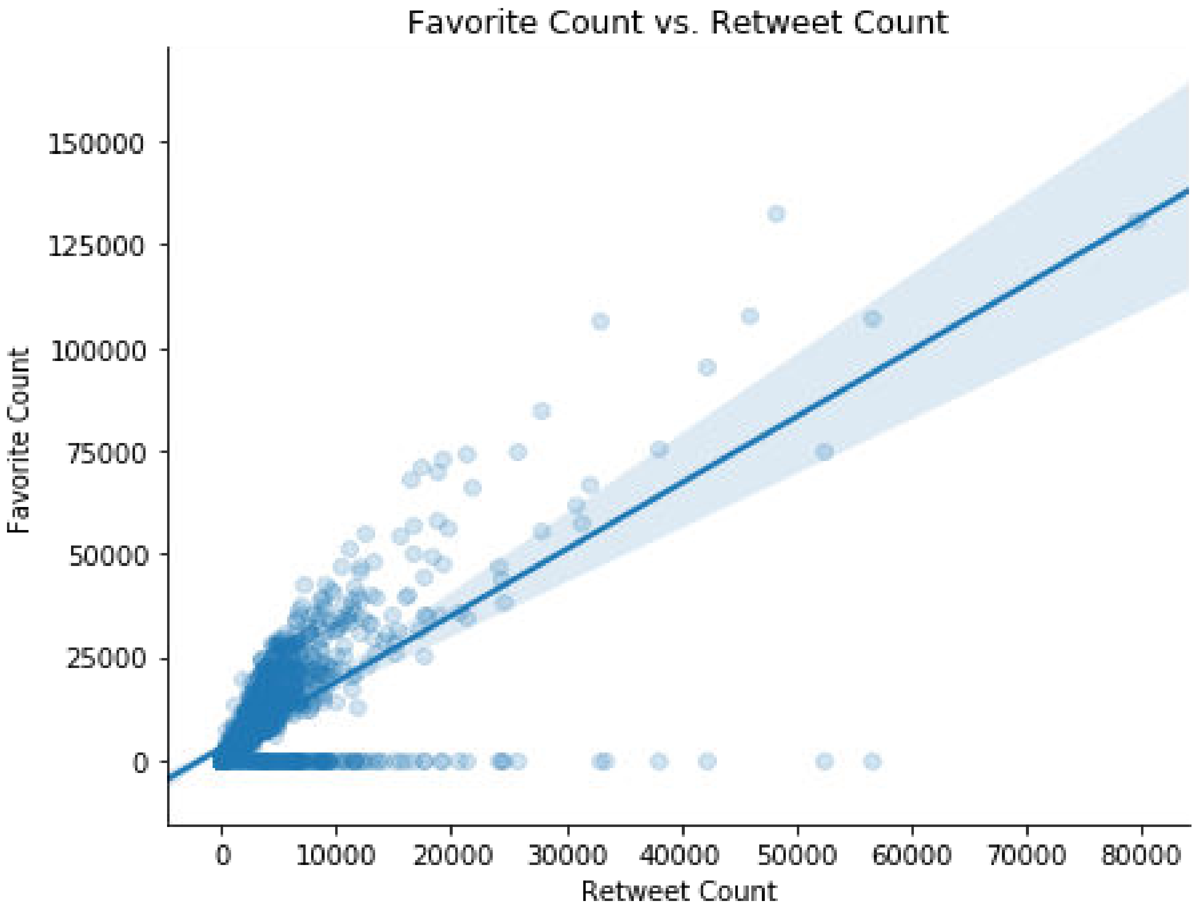
1- Visualizing the total number of tweets over time to see whether that number increases, or decreases, over time.

Over the time period of the tweets collected for this dataset, tweets decreased sharply starting in early 2016 (i.e. is 2016-01). While the tweets continue to decline over the time, there are spikes in activity during early 2016 (i.e. 2016-01) and in mid-summer of 2016, but continues to generally decreased from there. The owner of the WeRateDogs we must take this data into consideration to increase user traffic.



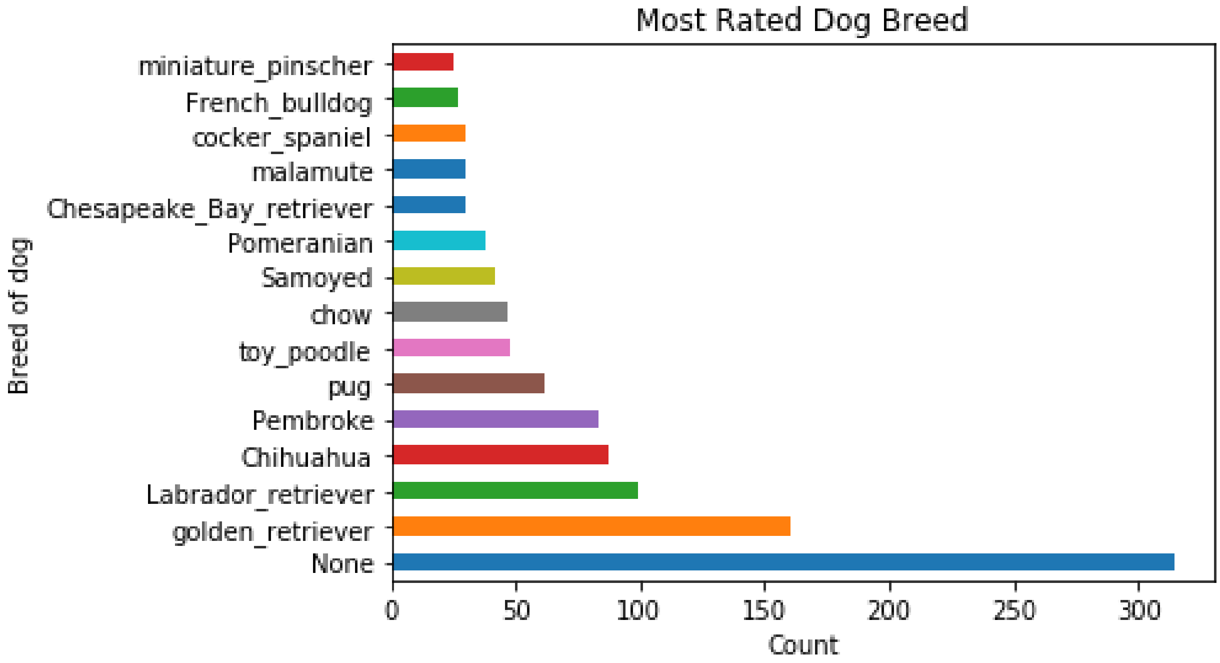
2- Visualizing the retweet counts, and favorite counts comparison over time.

There is a positive correlation between favorite (“like”) counts, and how much a post was retweeted. This correlation is important for the admin of the WeRateDogs twitter account to understand when determining method to increase users’ traffic on the page. we recommend previous posts with either a high retweet counts or high favorite count so that page admin could model future posts off historically popular posts.



3- What is the most popular dog breed?

The most popular dog breed is golden retriever. with a Labrador retriever coming in as the second most popular breed. Chihuahua isn't far bind. The page admin could use this information to create targeted marketing efforts for certain breed that aren't popular to increase their popularity, but also utilize the breed that are proven to be popular to drive user traffic to the page.



4- What is the most popular dog name?
The three most popular dog names are:

a) Lucy
b) Charlie
c) Oliver

Last but not least, we finished our report, there is a lot of visualization and analysis we can do as we have a huge data, but we mentioned here only sample of what we can do. as start and will expand our career with a lot of experience from what we gained at Udacity.

Thanks to all
Best regards,
Mohamed Fathy