Alex Paz
Wrangle and Analyze Data
Data Analyst NanoDegree
18 May 2019

WeRateDogs™ Twitter archive

This project was focused on gathering and cleaning data obtained from a number of sources, including CSV and TSV files, the Twitter API and JSON-formatted flat files. The original dataset had over 5,000 records of twitter posts by the twitter channel specialized in rating dogs in a fun and relaxed way: WeRateDogs (@dog_rates).

WeRateDogs™ (author)
@dog_rates

This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 puppared puppo #BarkWeek

1:02 PM - 24 Jul 2017

In order to gather, assess and clean the data programmatically and visually, Python programming language was our main tool, along with a few of its data-centered open-source libraries and some auxiliary ones (pandas, numpy, requests, matplotlib, and tweepy, to name a few). The main wrangling and cleaning file was created using Jupyter notebook, which allow for blocks of code, text and charts to be combined in one single document.

Information on the popularity of the tweets was acquired through the Twitter API, to which we had access after authenticating and obtaining the necessary developer approvals by Twitter.

When extracting data from one or more sources, it is considered essential to assess the quality and structure of the obtained data before any useful analysis can be performed. In this case, we noticed over 13 issues related to quality and two regarding structure (or *tidiness*).

As an example, the quality issues included inconsistency of data format, the presence of outliers and unidentified null values and variables that needed rearranging in a number of columns. All of the issues were duly addressed and tested through code, resulting in three clean tables which were then combined, in accordance with the widely-accepted *rules of tidy data*, so that we could extract initial insights and assess whether a more complete analysis should be performed.

“We noticed over 13 issues related to quality and two regarding structure (or tidiness)”

-Alex Paz

(images via WeRateDogs)

<u>Preliminary Insights:</u>

1. What are the most popular dog names on posts by WeRateDogs?

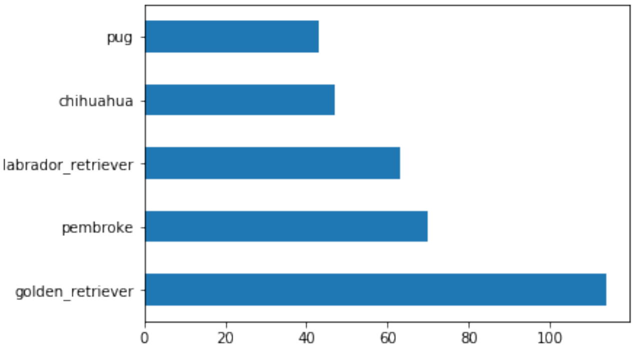
A: Most popular dog names were Lucy, Charlie, Oliver, Cooper, Tucker and Penny between Nov 2015 and Aug 2017.

2. What were the predicted breeds in the most popular tweets?

A: Labrador retriever, Eskimo dog and Chihuahua.

3. What were the most frequent breeds predicted by the neural network in this dataset?

A:


pug
chihuahua
labrador_retriever
pembroke
golden_retriever