Investigating a dataset for weratedogs

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

Using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. We will document our wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) . The dataset that we 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.

Gathering data

Data was obtained from three different sources:

(1) The WeRateDogs Twitter archive - WeRateDogs downloaded their Twitter archive and emailed it directly to Udacity so that we could use it for this project. This collection contains all 5000+ of their tweets' basic tweet information (tweet ID, timestamp, content, etc.) as of August 1, 2017. This file was manually downloaded

(2) The tweet image predictions - According to a neural network, this file (image predictions.tsv) is contained in each tweet. It is hosted on Udacity's servers and may be downloaded this URL :https://d17h27t6h515a5.cloudfront.net/topher/2 017/August/599fd2ad_image-predictions/imagepredictions.tsv .

(3) Additional data from the Twitter API - We collected the number of retweets and favorites ("likes") for each tweet. Using the tweet IDs from the WeRateDogs Twitter archive.

Assessing Data

We assessed the data and found the following issues;

Quality issues

1. Unnecessary data in the source column.
2. Removing retweet rows that have non-empty rows.
3. Irrelevant columns in the dataframe.
4. The name of the dog in some rows is indicated as an,not, a, one, the, quite, all.
5. Some row contains denominator rating of not equal to 10.
6. df1 primary key is represented by different name to the other datasets.
7. The indicators for missing values in the data varies NaN/None.
8. Erroneous datatype.

Tidiness issues

1. The columns doggo, puppo, pupper, floofer are in different columns.
2. Renaming the columns properly.

Cleaning data

We cleaned the issues above one after the other until all the issues were solved.

Storing data

After we were done with the cleaning we stored to a new dataset, where we used for the analysis and visualization parts.