Reporting: wragle_report

Project Overview

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.

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

There is no doubt that I learned many skills during this project. I really enjoyed making all the data wrangling process in this project and now I'm confident and capable to gather the data from any recourse and make an analysis on it. However, below I have explained all the data wrangling parts that will help you to go throw the code and understand it. Also, there is another document for communicating the result and visualizations act_report, I suggest you go and read it after reading this document.

Part 1: Data Gathering

Through this project, I did all the data wrangling process starting with gathering data using different methods such as reading the data manually and programmatically. In programmatic data gathering, I used web scraping technique to extract some data from Twitter API using Tweepy library and stored it in tweet-json.txt. After that, I used JSON library to read tweet-json.txt file line by line into a pandas DataFrame with these columns tweet ID, retweet count, and favorite count.

Part 2: Assessing Data

In the assessment part, I also used two methods to assess the data,

- **Visual assessment**
- **programmatic assessment.**

both are handy and useful in assessing data and both are used to identify the quality and tidiness Issues. In the quality Issues, we search for missing unreasonable, duplicated data or any column that has wrong data types, on the other hand, tidiness Issues is related to any structural Issues on the DataFrame, for example, two columns need to be split or two DataFrame need to be merged. In my case, I Identified 8 quality Issues and 3 tidiness Issues. The 8 qulaity Issues are as follows:

Tidiness issues

1. No need for four columns(doggo, floofer, pupper and puppo), make it one column that specifies the dog stage or type.

2. tweet-json table: no need for a separated table, It can be joined with Twitter archive in one table.

3. New column needed for dog breed

Quality Issues

twitter_archive table
There're 6 columns have missing data.

Wrong Data type in **timestamp** column.

Many column has the wrong data type **(retweeted_status_id, retweeted_status_timestamp, retweeted_status_user_id ,in_reply_to_status_id, in_reply_to_user_id)**

incompatible ID data type, **tweet_json.tweet_id** is string and the other two table are int.

Unreasonable name for 55 dogs named **a**.

some of the columns unnecessary such as: **retweeted_status_id, retweeted_status_user_id** and **retweeted_status_timestamp**,

Some Of the denominators isn't 10.

Wrong datatype for favorite_count and retweet_count columns.

Part 3: Data Cleaning

After Identifying all the Issues in the assessment part now we ready to data cleaning part and fix all the Issues that Identified above.

In the cleaning part, this is the most exciting part of the data wrangling project for me, you fix and clean all the issues using your skills and judgment. I started this part by copying all the three tables I have, to make the data cleaning on them, and the reason is that we want to save the original dataset from any modification or mistake. Then, I followed **define, code and test** approach to document every Issue Identified previously in the assessing part and write a code to fix it and finally test your code if it works and solves the problem or not.

Part 4-5: Storing, analysing and visualizing Data

Finally, I end up with one clean DataFrame after merging the three dataframe and clean them from all the Issues I Identified. I stored this DataFrame in a file called twitter_archive_master.csv. Now the data is ready for the analysis process, but before I start the analysis process, I must ask some questions that I want to answer after analyzing the data and exploring the answers. So I asked three questions:

1. What is the most common name for dogs?

2. What is the stage for dogs that get the most likes and retweets?

3. Which dog breed gets the highest rating?

To find the answer, please go to act_report document where you will find the answer and the visualizations for all the questions.