Wrangle Report

WE RATE DOGS

Antonius Nesseim | Data Analysis professional track | March 2021

Introduction

The purpose of this project is to practice what we learned in data wrangling course from Udacity Data Analysis Nanodegree program. The dataset that was wrangled is a 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. This report briefly describes wrangling efforts that had been done by me to finalize that project .

Data wrangling steps

1. Gathering data
2. Assessing data
3. Cleaning data

Project step by step

first Gathering data:

To Gather that project data, we need first to get 3 data sets form using three different ways :-

1- Twitter archive: the twitter_archive_enhanced.csv is a data set that provided by Udacity to be downloaded manually
2- Twitter Image prediction: That file “image_predictions.tsv” is hosted by Udacity server to be downloaded programmatically using the Requests library and URL information
3- Twitter API & JSON file: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing data

Once the three data sets were obtained, I started to assess the data as following:

- Visually by printing the three entire data frames separate in Jupiter Notebook.
- Programmatically by using python different methods (info, value_counts, sample, duplicated, etc).

VALUE_COUNTS, SAMPLE, DUPLICATED, GROUPBY, ETC

Then I separated the issues encountered in quality issues and tidiness issues. To prepare the points that I will work on it on the project next step.

cleaning data

first step was to create copy of the three original data frames to work on to avoid the original data frame damage

after making the new copy I was ready to proceed on cleaning the below issues on my new copy

1. Quality issues

Completeness, validity, accuracy, consistency (content issues)

A- twitter_archive

1. Keep original ratings (no retweets) that have images.
2. Delete columns that won't be used for analysis.
3. Erroneous datatypes (doggo, floofer, pupper and puppo columns).
4. Separate timestamp into day -month -year.
5. Correct numerators.
6. Correct denominators other that 10.
7. Correct the incorrect dogs names.

B- *image_prediction*

1. Drop the duplication.
2. Create one column for the prediction and other one for the confidence level.
3. Delete the unused columns.

C- tweet_json

1. Keep original tweets only

2 Tidiness issue

1. Change tweet_id to type int64 in order to merge with the other 2 tables " All tables should be part of one dataset"
2. combining the columns doggo, puppo, pupper, floofer into a single column

To clean data we had to took every issue and run the a code to solve that issue then test the data frame code result .

There were a couple of cleaning steps that were very challenging. One of them was in the image prediction table. I had to create a 'nested if' inside a function in order to capture the first true prediction of the type of dog. The original table had three predictions and confidence levels. I filtered this into one column for dog type and one column for confidence level.Other interesting cleaning code was to melt the dog stages in one column instead of four columns as original presented in twitter archive.

Conclusion

Programmatically Data wrangling is a very useful skill that any data analyzer should be familiar with .Using Python programing language and packaged make it some pretty easy to deal with data (and big data)as it is more powerful than using spreadsheets specially in the filed with big data . It is easy to document each single step and if needed re-run each single step. Thus, one can leave a perfect audit trail .