wrangle_report

June 28, 2022

1 WRANGLING REPORT

The data wrangling project was focused on wrangling data from @WeRateDogs Twitter Account. The dataset was gotten from three sources and in different format respectively. A summary of what I did during this Project is, I Queried Twitter's API for WeRateDogs Twitter Data such as Retweet count and Favorite count. Then combined these pieces of information with the additional data provided by Udacity Data Analyst Platform

Let's dive in, #### GATHER DATA FIRST I started off by importing packages I thought would be needed in this project such as numpy, pandas, requests, os, json , just to mention a few.

The first source is the Twitter Archive data. The Twitter Archive data was downloaded programmatically by Udacity and provided to me in a csv format o download manually. It was read into a pandas data frame using read_csv.

The second source is the Image Predictions data provided in a URL which I programmatically downloaded using the Python Requests library. I started off by creating a folder for the file then using requests to request for the file using URL, getting response 200 I knew was successful. I opened the file in the folder and employed 'wb'(write binary) while doing this because I'm dealing with Images. Then read the file into pandas data frame.

Due to some missing variables in the Twitter Archive table/data frame, I decided to query Twitter account WeRateDogs for additional information using an API. This is the third source.To begin with, I applied for a Twitter Developer account, got the necessary keys needed and got to work. I only needed more information from tweet ids I already had in Twitter Archive Data frame. So, I created a for loop for each tweet id in twitter archive to get the status. While using a for loop I made avoidance for errors using try and except. I also used a timer to check how long it took my code to run. It took about 30 minutes. I received 29 error messages telling me 'No Status found with the Id'. I created an empty list which I was going append. I wrote a code to open the Json file in read format and for each line loaded it to a variable. I accessed this variable to get the id, retweet_count and favourite count values. Then, I appended/ joined this to the empty list I initially created. I'm working with data frames so I convert this list to a data frame using pd.DataFrame.

Let's find dirt, #### BY ASSESSING DATA

I Assessed the data looking out for data quality tidiness issues. For Visual Assessment, the use of Excel spreadsheet came in handy, and I scrolled through the pandas Dataframe checking visually for any issues which I documented. Programmatically assessing the data was an easier process than visually. I assessed each dataframe individually using functions sch as head, tail, sample, info to check for nulls and incorrect datatypes. I noticed invalid data such as the name of a Dog being a, an, and the. Also noticed that the rules of tidy data were violated in the Twitter archive data, where dog stages (one variable) was represented in 4 columns violating the rule

'Each variable is a column'. Also, the retweets count, and favourite were on a different table than the Twitter Archive hence violating the third rule of tidy data which states that 'each Type of observational unit is a table'. I noted down all data quality and tidiness issues I spotted in the three dataframes then moved on to Data Cleaning.

Finally, #### CLEAN DATA. This stage had me getting my data from an Unstructured, untidy, and dirty format to structured and clean data. Based on the project motivation, I dropped rows of data where Tweets where not Original but Retweets. Removed, invalid data, Joined dog stages column into one. Merged the three data frames as one. I also dropped columns I thought to be redundant or not particularly useful for my analysis at this stage, Finally converting datatypes to the right format to aid my analysis.

SAVE After Cleaning, I saved my combined dataframe.