wrangle_report

July 27, 2022

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

INTRODUCTION

The objective of the study was to gather, clean and analyse over 5000+ tweets from a twitter account @dog_rates (WeRateDogs), draw some insights and make a visualization to communicate one of the insights drawn. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention both for its popularity and notoreity.

As always, the entire data wrangling process contains three major steps, namely:

Data gathering Data assessing Data cleaning Although, we did go beyond by both analysing the data and visualizing it with a plot. This was necessary as the data wrangling process isn't a stand alone process and for complete analysis, it must be combined with other processess to arrive at conclusions. I would be explaining in brief details the step-by-step prodecure that was undergone in this wrangling process.

DATA GATHERING

In this phase, three datasets were indicated to be necesary for the analysis.

One was provided beforehand (twitter_archive_enhanced.csv), and all that was needed to be done was to read it in as a CSV file into a dataframe using the pandas' read_csv function. The second file (image-predictions.tsv), was to be downloaded programmtically using the 'requests' library and then stored into a file on the pc before being read into a dataframe using the read_csv function again. This time though, it was required to specify the delimiter as an argument because tsv files are separated using tabs instead of commas like in a csv file. The third file was the most difficult to gather, partly because the concept was quite new to me and also because it required requesting access to Twitter's API. Unfortunately I wasn't granted access on time to the API so I opted to use the alternative that was provided by Udacity. This alternative involved downloading an already provided text file (tweet-json.text), and then reading the file line by line into a pandas dataframe. This brought me to the end of the gathering process with the three files successfully read into three different dataframes.

DATA ASSESSING

For this phase, I had to look at the three datasets both visually and programmatically to be able to notice and assess any possible quality or tidiness issue in the datasets before cleaning them in the next phase. A couple of issues were noted and documented in preparation for the cleaning phase. As with most data set, there were two majoy types of issues with the dataset, tidiniess and quality issues. The documented assessment is given below.

0.1.1 Quality issues

1.The name column has inconsistent data entry such as "a",'an","quite","none","the","space","old", in the twitter_archieve dataset

2.In the twitter_archieve dataset, "None" is an incorrect data entry in the doggo,floofer,pupper and puppo columns.

3.tweet id should be an string and not an interger to avoid calculation when the describe.

4.columns stated by the question to be dropped includes,'retweeted_status_id","retweeted_status_user_id",'re

5.The timestamps should be corrected to the appropriate datatype(datetime)set.

6.filter for only dog images.

7.Extract the source of the data from the surrounding url.

8.Some of the dog names in the 'p1', 'p2' and 'p3' columns aren't consistent,some start with upper case while the other are all in lowercase.

0.1.2 Tidiness issues

1.Merging of the doggo,floofer,pupper, and puppo column as one single column.

2.The three datasets should be merged into one dataset as all the three datasets are part of the same observational unit

DATA CLEANING

Data cleaning basically involves tackling the quality and tidiness issues that were note in the previous phase - Data assessing. With the use of formulas, functions and loops, most of the issues were cleaned. As was required by Udacity, the entire datasets were not to be cleaned in it's entirety. Rather, a minimum of 8 quality issues and 2 tidiness issues were to be worked on.

During the cleaning the datasets, I combined the three into one master dataset (twitter_archive_master.csv) and saved it in preparation for analysis and visualization.

For insights on analysis and visualization, please refer to the 'act_report.html' file which contains a brief summary of the insights gotten from analysis.

In [ ]: