wrangle_report

September 25, 2022

0.1 Reporting: wrangle_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.

0.1.1 WeRateDogs Twitter Project

In this project,and specifically explaining the wrangling effort, the data goes through three important stages which include gathering, assessing and cleaning.

The gathering stage involve asking and answering the question of which data are needed to provide answers to the problem statement or solve a problem. In this stage, relevant python libraries like *pandas*,*numpy*, *tweepy*,*request*, among others were imported into the jupyter notebook to aid this process. Data for this project is acquired from three different sources. One data source is a csv named twitter_archive_enhanced.csv, that has been provided already. This data was read into a dataframe named *Twit_archv*. Another data source was scraped from Twitter using approved credentials and Application Programming Interface (API), the with function and json.load methods were used to read the file into a dataframe which was named *tweet_data*. The last data source was acquired from the web using the url hosting it on the web. The request function in pandas was intrumental to culling this data from the target url. A folder was created to hold the downloaded file from where it was read into a dataframe using read_csv function, taking cognisance of the file path. This file was named *image_pred*.

The next phase of Data wrangling for the project is assessing the data. In this stage, the three tables were checked for tidyness and quality issues. The tables were inspected visually where errors were spotted in `Twit_archive_Data` that the expanded_urls feature had unrelated urls and the feature for dog name in `Twit_archive_Data` table had some names captured as 'None". Also, the three tables were assessed programatically with some python functions like `describe`,`info` and `sample`which also revealed quality issues like missing data and wrong datatypes ascribed to some columns. All identified errors in the data are documented in a short and simple phrase.

The last stage of the wrangling process has to do with cleaning of the identified issue that include quality and tidyness problems. Retweet status and in-status reply columns were dropped from the dataframe as the size of available data were very small and inconsequencial. Other columns like expanded_urls with some missing data were corrected by dropping the rows with missing data in the expanded_urls feature. The rating_denominator is supposed to have a figure of 10, so rows beearing the complement of 10 were eliminated. Similarly, for rating numerator, the entries were plotted in a boxplot from where it was discovered that the data is skewed to the right with some outliers, so the affected rows were eliminated. Timestamp column was converted from object datatype to datetime format. The dog breedas found in p1, p2 and p3 were alsostandardized

to appear in sentence case. Finally, all three tables were merged on twwet_id using inner join to form a consolidated dataframe for the weratedogs project. This dataframe was saved in csv format with file title *tweet_archive_master*.