wrangle_report

February 12, 2023

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.

0.1.1 INTRODUCTION

This report seeks to breifly describe the process of removing errors and combining complex data sets to make them more accessible and easier to analyze. Three different datasets were wrangled, multiple quality and tidiness issues identified and each steps will be discussed in sections below. Visual and programmatic assessement was used assess the data, several iteration of the data wrangling process was also employed to ensure a clean resulting master dataset.

0.1.2 Quality issues

Twitter archive table: Practically empty/missing data across multiple columns (retweeted_status_id - retweeted_status_timestamps,in_reply_to_user_id, in_reply_to_status_id) Some columns in the Twitter archive table datasets were quite scanty - less than 10% of the row are filled with usable data. One option is to fill up the empty spaces with for example average value of available data but this option is not applicable in this case. Attempting to fill up will be commiting data fraud as each field is expected to be unique and verifiable - for example "retweeted_status" for a particular tweet.

Missing Data in columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'

There is a possibility that the Columns 'doggo', 'floofer', 'pupper', 'puppo', are dog names and unique columns were created erroneously. This will be true if each unique entry has none value in the corressponding 'name' column. This ascertion is confirmed as false using command: df_twiter_archive[df_twiter_archive['doggo']=='doggo'].head() for each column name.

Data in listed in all columns above are less than 10% which quite negligible and was dropped mostly because we have no supplimentary data.

Twitter archive table: invalid or missing dog name - name represented with single letter or 'None' Dog names were represented with single characters or impossible words like 'the'. All occurences of such single character was replaced with 'none' for uniformity. its better to have no dog name inputted than those that are likely erroneous.

Twitter archive table: ’timestamps’ column is in object format Twitter archive table: ’timestamps’ column was observed to be in object format. Data manipulation is easier when columns are in the right timestamps format. This was programtically resolved.

Twitter archive table: Remove html tags from URLs in source column HTML tags present in this column makes it impossible to directly lift the data and consume programatically. This contamination is remove with the code below.

Twitter archive table: Remove retweets Retweets are removed from the Twitter Archive table as part of the project instructions.

Twitter archive table: Remove ratings and URL in text column Present of data like URL and ratings in the text column violate primary data wrangling rule. This was programatically cleaned.

Twiter archive table: Null values in ’expanded_urls’ column This exercise was simply to remove empty rows in the "expanded_urls" column in the Twitter Archive table

Image_prediction table: Invalid p1,p2,and p3 columns dog names such as hen, paper_towel etc A number of invalid names for Dogs will removed. List of invalid words is first declared then replaced with the word ’none. Invalid name names including ’cup’, ’studio_couch’,’sliding_door’, ’minibus’, ’toyshop’, ’bald_eagle’,’binoculars’ and more.

0.1.3 Tidiness issues

Image-prediction file column names are not descriptive In the Image preiction table, a number of column are not descriptive. Column names like ’p1’, ’p1_conf’, ’p1_dog’ was changed to ’prediction1’, ’prediction1_confidence’, and ’prediction1_validity’ respectively.

df_Tweet table should be part of Twitter archive table As part of wrangling activities, df_tweet table was identified as better part of the Twitter Archive table instead of being a stand alone table with only three columns

```
In [ ]:
```