wrangle_report by Ayomide Adenuga

February 22, 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.

0.2 Wrangle report By Adenuga Ayomide

Introduction: This wrangle report is part of the Wrangle and Analyze data project by Ayomide, the data used in this project is the twitter archive data from a twitter handle @dog_rates. This handle is also known as WeRateDogs, WeRateRogs is popular for rating dogs in humorously. The wrangle exercise is in three phases, viz: gathering data, assessing data and cleaning data.

Gathering I gathered data from different sources and different format in this project. - A. The WeRateDogs twitter archive enhanced dataset was provided by Udacity, the data is in comma separated value format(CSV). All I did was to download it directly from Udacity website.

- B. The tweet image predictions. The file is download programmatically with the python request library, and it is hosted on Udacity server.
- C. The tweet json file was provided by Udacity, I should have query Twitter API, but I my twitter account does not have elevated status and there is no longer time by my side.

Assessing The data was assessed both programmatically and visually, this is done in order to discover both tidiness and quality issues. The data failed the quality issues test if it does not pass the following: each variable forms a column, each observation forms a row, each type of observational unit forms a table. Quality issues Twitter Archive Enhanced Table (df_1) Table 1. timestamp data type is string and remove +0000 2. rating_denominator column contains inaccurate data 3. rating_numerator column contains inaccurate data 4. Erroneous datatypes (tweet_id) 5. In the name column, None is used to represent some dog name 6. Text in source column is not readable

0.2.1 Image predictions (df_2) Table

1. Dog name in P1, P2, and P3 is separated by underscore instead of space and start with lower case letters
2. Erroneous datatypes (tweet_id)

0.2.2 tweet json (df_3) Table

Erroneous datatypes (created_at, id) Tidiness issues 1. Column header in df_1 (doggo, floofer, pupper, puppo) should be in the same cloumn 2. The three tables should be combined into one. 3. There are duplicated columns and columns that won't be used in the analysis

0.2.3 Cleaning data

The quality and tidiness issue identified in the cleaning phase guided the cleaning phase, all the quality and tidiness issues documented were addressed, and this was done in three phases, define, code and test. The define section explain the issue and what will be done to correct it, the code section is about the code that was written to clean the data, while the test section is to ascertain whether the code written actually corrected the issue.

```
In [ ]:
```