wrangle_report

January 11, 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.

1 Wrangle Report

1.1 Introduction:

The wranlge report is part of the Data Wrangling project which consists of gathering data from a twitter account also known as WeRateDogs. This account rate dogs based on comments about dogs and this project allowed me to used the data analytics process by gathering, assesing, and cleaning the data before using visual tools to illustrate the findings.

1.2 1. Gather

We used several souces in this project including:

- WeRateDogs Twitter archive that was provided by the Udacity website and downloaded through there.
- I used the Requests Library to download the JSON file and was able to get the count of retweets and favorite counts.
- Image prediction file was also included in Udacity's website and I was able to download it through the Requests library.

1.3 2. Assesing the data

After downloading and assesing thre data correctly, my next step was to identify the quality and tidiness of the data at hand. The quality of the data indicates how accurate the data is in relation to the content and tidiness refers to how readable or visually the data presents itself. To do this I used several python codes to identify errors in the data, including: duplicated, query, value_counts, and describe. For better understanding and organization, I included the findings in each section of every file for the purpose of targeting those areas in the cleaning process.

1.4 3. Issues Found:

1.4.1 Quality issues

- Consistency of the data
- The accuracy in which it is presented
- Completeness
- Duplicate Data
- Ambiguous or misleading

1.4.2 Tidiness issues

- Every cell has a single value
- Every column is a variable

Observations (twitter archive)

- Data is incomplete containing plenty of null values
- There are several incorrect data types, for example: in_reply_to_user_id and retweeted_status_id are among data types that shoulbe be swtiched from float to int.
- The rating numerator contains abnormal values including large numbers such as 1776,960.. etc
- Timestamp data type should be datetime

Observations (image prediction)

- Many image predictions are not dogs even though the image contains a dog in it, for example: desktop computer, suit, bow, etc...
- There are duplicated image urls in the data

Observation (json table)

- Need to be joined with archive table

1.5 4. Cleaning process

Some of the issues were addressed and cleaned during the process. It would probably take a long time to identify and clean the entire dataset as more issues probably exist; however, for our visual aspects or my intention; the data is clean enough to make my visualizations. In the process I needed to create a clean version of the original files I was working with; this way we kept the original file in case we needed for later use. In the cleaning process I was able to remove the null values, remove certain columns I was not going to be working with, change the data type for timestamp to datetime, remove duplicated values, and join the json table with the archive table. These fixes allowed me to illustrate or present the data in a clear and more precise manner.

```
In [ ]:
```