wrangle_report

March 19, 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 DATA WRANGLING REPORT

1.0.1 WeRate Dogs Data

Data Gathering phase:

```
The inception of my project began with manually downloading and uploading the twitter-archive-en
into my juypyter notebook workspace.The next dataset, image-predictions.tsv was programmatically
link provided by Udacity, which was uploaded and read into a pandas dataframe in my jupyter note
I had setback acessing the twitter-api, so I manually downloaded and uploaded the 'tweet-json.tx
Each data uploaded was read into different datafrmes with distinct table names for easy identifi
```

Assessing And Cleaning Data phase:

```
Some quality and tidiness issues were identified for the three tables. Detaiils of the assessing
```

Quality

```
Corrections were made to some Quality issues encountered during assessment and cleaning.
```

Tweet archive enhanced (archive_df)

```
-Too much missing data from some columns in this table. The columns [retweeted_status_id, retwee
-The issue of rating_denominator column value exceeding 10
-Errorneaus column label,column labeled 'text' when we dealing with tweets.
-Incorrect names like 'a', 'the','an', in the ['name'] column. Those are not actual names of peo
-'None' values in the ['name'] column instead of NaN to fill in rows with missing values is an i
```

Image Prediction Table (df_image)

```
-The columns, [p1,p1_conf, p2,p2_conf, p3, p3_conf...] cannot be understood in the table, they c
-Delete duplicate values
-Dropping unwanted column ['img_num'] column
```

Tweet_json (df_tweet_json)

-Errorneaus column "id" since we working with tweet ID's in that column

Tidiness issues

-Merge all three datasets into one dataset, 'archive_clean', df_images and 'tweet_clean' data.
-Merging all types of dogs into a single column 'dog_stage'

In [ ]:

In [ ]: