wrangle_report

September 12, 2022

0.1 Reporting: wragle_report

0.1.1 Data Wrangling

Data wrangling steps include Gathering of data, assessing of the data and Cleaning of the data. #### 1. Gather First of all, the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv) was directly downloaded and then imported into the workspace using the pd.read_csv command. The Requests library was then to download the tweet image prediction (image_predictions.tsv) using requests.get(url) command. The last part of the data gathering step was using the Tweepy library to query additional data via the Twitter API (tweet_json.txt).

2. Assess Nine (9) quality issues and two (2) tidiness issue were detected and documented. They are as follows: ##### Quality Issues 1. There are duplicate images for some entries. 2. Some column names are confusing as they do not give much information about the content. 3. There is more than one classification for some dogs. 4. Some dogs have no classification. 5. The image prediction dataset has 2075 entries as compared to twitter archive's 2356 entries. 6. There are some missing values. 7. The number of data entries in the twitter API dataset differs from the other two datasets. 8. There are some wrong dog names. Eg. a dog being called 'a' 9. Timestamp has wrong datatype that is object instead of date. ##### Tidiness Issues 1. All the datasets should be merged. 2. The columns that classify dogs should be added together for easier analysis.

3. Clean Some of the issues documented were accessed. These issues are as follows: 1. There is more than one classification for some dogs. 2. Timestamp has wrong datatype that is object instead of date. 3. There are duplicate images for some entries. 4. There are some missing values. 5. Some columns are not necessary 6. New column needed. 7. Some column names are confusing as they do not give much information about the content. 8. Reordering column names.

In [ ]: