wrangle_report

October 4, 2022

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.

Data were gathered from different sources, assessed, cleaned and visualsed for insights. Below are three insights gathered from the data. ### Sources of data: The data were gathered from three sources: 1. twitter_archive(twitter-archive-enhanced.csv) 2. image prediction(https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) 3. Twitter API.

0.1.1 Data assessment:

A number of issues were identified for assessment. They include:

0.1.2 Quality issues

1. Incorrect dog name
2. Remove retweet rows
3. Drop the columns not needed
4. Incorrect tweet_id datatype
5. Incorrect datetime format for timestamp
6. Inaccurate datatype for rating_numerator and rating_denominator
7. Inconsistent dog breed names capitalisation
8. Drop duplicate values from jpg_url
9. Replace the urls in the source column by the source name

0.1.3 Tidiness issues

1. Merge all three datasets together
2. All dog types should be in the same column
3. Create a column for dog breeds

0.1.4 Cleaning Data

After identifying the quality and tidiness issues stated above, the gathered data were cleaned, first by defining what the issues are followed by writing the code to clean the data and lastly testing to see that the data is cleaned.

0.1.5 Storing data

After cleaning the data, it is then stored as twitter_archive_master.csv. This was the final stage before the analysis and visualisation of the data