wrangle_report

January 14, 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.

WeRateDogs is a twitter account that allows their followers to rate dogs by favoriting and retweeting them. A twitter API was used to query information from the WeRatDogs account of which an access to the data was given for our analysis. The wrangling process entailed Gathering, Assessment and Cleaning the data. The data was gathered by directly downloading the we_rate_dogs_twitter_archive.csv file, using the requests library to download the image_prediction.tsv file and finally tweepy to query additional data from Twitter. the three diffrent tables were then combined to one single dataset for easy assessment.

in Assesing the data, a copy of the main file was made and used for the analysis. several Quality and Tidiness issues were identified, first visually and then programmatically. The identified quality issues entailed some columns that had very low data and therefore would not help our analysis. These columns were dropped during the cleaning process. some dog names were spelt wrongly, others with Capital letters and others small letters. for consistency, all were chnaged to small letters and the wrongly spelt names corrected. some missing values were not well documented and the text column contained unnecesary additions all these were corrected after cleaning. retweets and replies were also removed keeping original tweets and all data types were correctly changed as some had wrong data types. timestamp was also converted to Datetime to have a correct date and time for each tweet. Some tidiness issues had to do with dog stages split into four (4) different columns making it repeated and also expanded url column had some rows with multiple urls that was not so relevant to our analysis. After these issues were documented, they were now cleaned accordingly to have a much better dataset for easy analysis.

After cleaning the data was stored to a csv account and then downloaded again. Some insights were generated from the final dataset which included the most popular dog names, the most common tweet source and the most favorited and retweeted tweet. Most popular dog names were found to be cooper, charlie, lucy and oliver and the most common tweet source was the Iphone. the most common dog names and tweet sources were visualised using matplotlib.

In [ ]: