# wrangle_report

May 22, 2022

## 0.1 Reporting: wrangle_report

### 0.1.1 Gather

Here, the three dataframes are required to be downloaded. The twitter_archive_enhanced.csv is downloaded manually by uploading it in jupyter notebook and read it directing into the pandas dataframe, while the tweet image prediction, Tweet JSON are downloaded programmatically. The image_prediction.tsv is using downloaded through the url using request library while, Json data the twitter api is queried and it is loaded into tweet-json.txt. ### Assess I assessed the three dataframes visually using '.head()' to have a scan through it, and programmatically using '.info()', '.sample()', '.duplicated().sum()', '.isnull().sum()', '.describe()' for a more in depth assessment for each of the dataframes. #### Copy of dataframes Before the dataframes where cleaned, a copy was made of each of them. ### Clean 1. I replaced underscore with space in 'p1', 'p2', and 'p3' 2. I capitalised the dog names in the 'name' column in archive_clean tables and the dog breeds in 'p1', 'p2', and 'p3' in image_clean 3. I changed the datatype for the'tweet_id' column in archive_clean and image_clean, the datatype for 'id' column in json_clean tables to string, and the timestamp datatype in archive_clean to datetime 4. I checked that there are no tweets after August 2017 and there was none 5. I renamed the 'floofer' column to 'floof' in the archive_clean table and 'id' column to 'tweet_id' in the json_clean table 6. I created dog_stage column for the different dog stages which are; doggo, floofer, pupper, puppo, and doggo. The columns were removed, it should be an observation on a row. 7. I corrected the rating_numerator that was incorrectly extracted. 8. I merged both the archive_clean and json_clean tables into one table first, then I merged the newly merge_df1 table and combine with the image_clean table 9. Dropping the retweets, reply records, we only need original tweets so I will drop the columns we do not need and drop the tweets that do not have an image or have images but do not display dogs #### Conclusion After doing all these, the gathered, assessed, and cleaned master dataset was saved to a CSV file named "twitter_archive_master.csv".

In [ ]: