wrangle_report

January 15, 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.

0.1.1 Gathering:

basicaly it's the first step in wrangling, following the instructions I gathered the data using twitter api, requests and the usual download manner. ### Assessment and cleaning: The 3 data sets we got contains a lot of quality and tidiness issues, so I did my best to clean it up. In Twitter archive data set there are missing data like in name, doggo floofer ... but its not NaN type. and this might be missleading sometimes so I replaced the value 'None' with NaN so that the pandas funstions such as .isna() and .info() can detect that missing values exist. The Twitter archive contains also retweet rows and these rows we don't need them, we only need tweets, so I dropped the retweet rows using the retweet columns (when then retweet columns contain no NaN value means it's a retweet) after that I deleted the retweet columns because we don't need them anymore. Twitter archive contains a tidiness issue (the columns doggo, floofer, pupper and puppo are variables) so I created new column called dog_noun that contains the dog noun it must be doggo, floofer, pupper or puppo, most of the time it contains NaN because it has a lot of missing values. Image predictions data set contains also some issues,it contains the data of three prediction algorithms and sometimes these algorithms says that the image does nt ontain a dog, So we had to delete the rows that the three algorithms do not detect a dog in their images. We needed also to add a new column that contains the final or the usable predicted result from the three algorithms (column name is pf). We faced also a completeness issue because the three data sets do not have the same count of rows so we had to join the three into one data set.

```
In [ ]:
```