# wrangle_report

May 22, 2023

### 0.1 Reporting: wrangle_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.

Introduction: The purpose of this report is to document the process of data wrangling performed on the WeRateDogs Twitter data. The data was gathered from various sources, including a Twitter archive, image predictions, and additional tweet data obtained through the Twitter API. The goal was to assess the quality and tidiness of the data and address any issues to prepare it for analysis.

Gathering Data: The data was gathered from three different sources. The main dataset, twitter_archive_enhanced.csv, was provided as a CSV file and was imported into a pandas DataFrame using the pd.read_csv() function. The image predictions data, stored in the image-predictions.tsv file, was downloaded programmatically using the Requests library and then loaded into a DataFrame using pd.read_csv() with the appropriate separator specified. Additionally, tweet data such as retweet count and favorite count was obtained by querying the Twitter API for each tweet ID in the Twitter archive and saving the JSON response in the tweet_json.txt file.

Assessing Data: The gathered data was assessed visually and programmatically to identify quality and tidiness issues. The assessment revealed several issues, including missing values, inconsistent data types, incorrect dog names, and inaccurate ratings. Additionally, the datasets were not structured in a tidy format, with multiple columns representing the same variable. The assessment findings were recorded and served as a basis for the cleaning process.

Cleaning Data: The cleaning process involved addressing the quality and tidiness issues identified during the assessment phase. Missing values were handled by either imputing or dropping the respective rows or columns. Inconsistent data types were corrected, and incorrect dog names were replaced or corrected based on further inspection. The inaccurate ratings were extracted correctly from the tweet text using regular expressions. To improve tidiness, the dog stage (doggo, floofer, pupper, puppo) was transformed into a single column instead of four separate columns.

Conclusion: In conclusion, the data wrangling process involved gathering data from multiple sources, assessing the data for quality and tidiness issues, and cleaning the data to address the identified issues. The resulting cleaned dataset is now ready for further analysis and exploration. The cleaned dataset, twitter_archive_clean, along with the cleaned image predictions and tweet data, provide a reliable foundation for gaining insights into the WeRateDogs Twitter data.