**WRANGLE REPORT**

**INTRODUCTION**

This project involved gathering, assessing, cleaning and analyzing data from tweets of “WeRateDogs” twitter account

Python packages used includes numpy, pandas, requests, os, matplotlib

**DATA GATHERING**

Data was gathered from three sources and read into dataframes for further assessment and cleaning

* The file named ’twitter\_archive\_enhanced.csv’ contained the archived data of tweets and was read into ‘twitter\_archive’ dataframe
* The ‘image\_predictions.tsv’ file contained data on image predictions of the dogs from the tweets. It was extracted from the given URL into a folder and then read into ‘image\_pred’ dataframe
* The given ‘tweet-json.txt’ file contained additional columns on retweet counts and favourite counts. It was read into a ‘tweet’ dataframe

**ASSESSING DATA**  
Visual and programmatic assessments of the 3 dataframes were carried out to identify quality and tidiness issues. The following issues were identified:

**Quality Issues:**

* Datatype issues: tweet\_id column is stored as integer in 'twitter\_archive', 'image\_prediction' and 'tweet' dataframes also timestamp column in 'twitter\_archive' is stored as string datatype
* Wrong dog names, the wrong dog names are noted to all be in lowercase
* Rows with replies and rows with retweets
* Irrelevant columns: 'in\_reply\_to\_status\_id', 'in\_reply\_to\_user\_id','retweeted\_status\_id' , 'retweeted\_status\_user\_id' and 'retweeted\_status\_timestamp' contain replies and retweet and are not relevant to analysis
* Some rating\_denominators are not 10
* Unclear column names description in image\_pred dataframe
* Some dog rating\_numerators were wrongly extracted
* From image\_pred dataframe, the three image predictions (p1\_dog, p2\_dog and p3\_dog) of some tweets were not dogs

**Tidiness Issues:**

* Dog stage variable in ‘twitter\_archive’ stored in four columns
* All the data should be on one table

**CLEANING DATA**

* To clean the dataframes, original copies were first made to keep the data intact
* Identified issues were then cleaned using the ‘Define-Code-Test’ Sequence

**SAVING DATA**

The cleaned data was saved as a master dataset. This was then saved to a CSV file named ‘twitter\_archive\_master.csv’

**ANALYSIS AND VISUALISATIONS**

Analysis and visualisations of the data was then done using pandas and matplotlib libraries

**INSIGHTS**

Insights gathered:

* The highest value of rating is 14
* Only 336 of the 2356 tweets had dog\_stages
* Of the tweets that had dog stages, pupper was the most, consisting about two thirds of the total
* Of the tweets with dog names, Obie's tweet seemed to have the highest number of retweets and favorites count
* For the image predictions, about three-fourths of the first predictions of images turned out to be dogs
* Out of the correct first predictions, 'golden\_retriever' is the most predicted dog breed