WRANGLING REPORT

The wrangling process started with downloading the required files i.e twitter archive and image predictions, save and converted to pandas' data frame for assessment, cleaning and analysis, then we added additional columns by scraping favorite count and retweet count using the respective tweet id from the twitter archive dataset. This process covers the data gathering process of the wrangling.

The next process encompasses both programmatic and visual inspection of the different data frame to spot out quality and tidiness issues to enable us come up with plans to arrest the different issues.

Quality issues are:

- Name column in Twitter_archive_df data frame contains some invalid records example "a"
- Columns with high amount of null in Twitter_archive_df data frame which will be categorize as low-level information columns
- ID Columns in the three data frame are in int format instead of string
- Timestamp column in the Twitter_archive_df data frame is in object format instead of datetime format
- Missing values in expanded URLs ought to be populated with a string representing missing values
- The Source columns needs to cleaned to present a more presentable value
- Name column in Twitter_archive_df data frame need to be changed to dog name as this new column name is better for information purpose
- Drop the second and third likely prediction.

Tidiness issues

- Each Variable does not form a column, as the various dog stages can be populated in one column call dog stage
- Merging will have to be carried out to attain the structure goal of only ratings with images