

## **ACTION REPORT**

The actions carried out to meet the required table specification and effectively clean the data as part of the wrangling process include:

- Cleaning the name column to remove and correct incorrect values
- Dropped columns with high number of null values and reflect low level information column
- Converted ID columns from integer to string
- Converted Timestamp column in the Twitter\_archive\_df from object format to datetime format
- Cleaned the Source column to a more presentable value
- Rename name column in twitter archive to dog name
- Dropped the second and third likely prediction
- Each Variable does not form a column, Populating the various dog stages in one column call dog stage to ensure each variable form a column
- Merged the data frame to attain the structure goal of only ratings with images

After this cleaning has been carried out, I proceeded to storage where I stored the merged file in csv format using pandas, then later moved to insights and visualization.

The few insights gotten from the data wrangling process include:

1. The maximum Favorite count and Retweet count was recorded for dog stage doggo with an unknown dog name and dog breed prediction of Labrador retriever with a favorite count and retweet count of 145822.0 and 71305 respectively.
2. 1532 predictions were dog breeds while the remaining 543 were other items from the image.
3. The source for the twitter archive information was majorly from twitter for Iphone with 2034 observation then twitter for web client with 30 observation and lastly twitter for Desktop with 11 observations.