Wrangle report

After gathering the three data sets: twitter_enhanced, image predictions and tweet-jason the assessment process was carried out to enhance the quality and tidiness of the data. All three data sets were assessed programmatically and visually. Nine quality and 2 tidiness issues were detected. The quality issues analysed in this report are listed below:

Quality issues

1. ID Columns in the three dataframes are in int format instead of string
2. Name column in Twitter_archive_df dataframe contains some invalid records example "a"
3. Some rows were retweets and need to be cleaned so that only original ratings are analyzed
4. Columns with high amount of null in Twitter_archive_df dataframe which will be categorize as low level information columns
5. Time stamp column is in object format instead of a date time object
6. The Source columns needs to cleaned to present a more presentable values
7. The second and thirdly likely predictions need to be dropped
8. Misssing expanded_urls in tweet enhanced data frame ought to be populated with a string representing missing
9. Extracting Ratings correctly

The first quality issue was to change the int format of id columns in three datasets to string format. Secondly the invalid names thus names which does not make sense such as a, no and officially were changed to unknown names. A list of invalid names was created and then all the names in this list were renamed to Unknown. Also names with lower case were put in the same category of invalid names.
Another major quality issue was to rectify incorrect numerator rating. It appeared that numerators that contained decimals were not extracted correctly from the source text. In the wrangling process the rows with decimal numerator rating were identified and the numerator was extracted correctly

Tidiness issues

1. For the dog stages each Variable does not form a column, as the various dog stages can be populated in one column called dog_class
2. Merging to have only one data set with ratings and image predictions

To make the data set tidy all the dog stages were condensed into one column called dog_class After corrected the quality issues all three data sets were merged into one csv data set twitter_archive_master to make it easier for visualisation and analysis purposes.

Lastly all three datasets: twitter_enhancd, twitter_jason and image_predictions were merged into one data frame to make the data tidy and easy to analyse and make visualizations.