Wrangle and analyze data

 Data wrangling, the goal of this project is wrangle and analyze data, to create interesting and trustworthy analyses and visualizations.

To wrangle data there are three steps should work with:

Gathering data, assessing data and Cleaning data.

Gathering data is one of the most important skills that any data analyst and data scientist working with.

In this project I gathered data from three different sources the first data frame gathered by reading ‘**Enhanced Twitter Archive’** is CSV file which contains basic tweet data for all 5000+ of their tweets, The second data frame gathered from TSV file **‘Image Predictions File’** by read this file form URL, and The last data frame gathered from Twitter.

All these data frames are about dogs, this is a twitter account ([@weRateDogs](https://twitter.com/dog_rates)) working to rate dogs by special rating system to rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

Assessing data, in this step I worked to investigate and explore the data frame to find issues. These data frames contain many issues some of them are Quality issues which is accuracy, validity, consistency, completeness and other is Tidiness issues which is structural issues.

Quality issues:

* Timestamp in twitter\_archive\_df is object, change it to datetime type.
* Text (tweets) contains #, extract the hashtags from the text column and store it in tweet\_hashtag column.
* Name have some names with (a, an, not..) which is not names.
* Calculate denominator and nominator to rating\_number.
* Delete the columns that I don't use it.
* Sperate timestamp into three columns, (day, month, year).
* Remove any denominator does not equal 10 (1,50,80,20,2,16,40,70,15,90,110,120, 130,150,7,0).
* Remove outliers from nominator column.
* Calculate the rating in new column and store the percentage value.
* Remove all retweets from dog\_images\_df and keep tweets.
* Some tweets without images because the number of rows in twitter\_archive\_df and dog\_images\_df does not match.
* None values to NAN in Name, doggo, floofer, pupper and puppo.
* Combine puppo, floofer, doggo, and pupper into one column 'dog\_stages'
* Remove duplicated values in twitter\_archive\_df.

Tidiness Issues:

* Merge all data frames together
* jpg\_url contains duplicated URLs.
* There are four different columns of dog stages (Doggo, flooder, popper and puppy) merge it into one column and remove nulls.

After Assessing data, should cleaning data frames to analyze it.

In this step,