Data Wrangling Report

This report will describe my wrangling efforts.

First, we define Data Wrangling steps.

There are 3 steps we will talk about each of them in detail, the three steps are:

1- Gathering Data
2- Assessing Data
3- Cleaning Data

The first step is Gathering Data.

Gathering Data

Data for this project came from three different sources:

- Twitter archive: This data given from Udacity team.
- Image predictions: This data given from Udacity team and we downloaded it programmatically.
- Twitter data: This data obtained from Twitter api.

Assessing Data

After we obtained these data we go to the second step (Assessing).

Assessing can be visual or programmatically.

First, I opened these three data to make a visual assessment and obtained these observations:

- Source column isn’t seemed good.
- Text column contains the text, rate and short_url.
- There are missing values in name column and there exist messy data.
- There are many missing values in (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_url) columns
- There exist unnecessary columns.
- Dog stage exists in 4 columns (doggo, floofer, pupper, puppo).
- p1, p2 and p3 columns contain uppercase and lowercase letters.
- id_str should be replaced into tweet_id.

All these observation I noticed it visually.

When I used the programming, I noticed all previous observation in addition to some other observations:

- (tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id) columns is not in appropriate type.
- timestamp, retweeted_status_timestamp columns should be datetime type.
- There are some mistakes in values for rating_numerator and rating_denominator.
- expanded_urls column got duplicates and links to different websites.

- tweet_id should be string.
- id_str should be string.

All these problems are solved in the next step which is cleaning.

Cleaning Data

Is the third step in Data Wrangling Process, and all previous problems can be solved in this step.

- Source column was written in html format, so I extracted the text I need from this format.
- I separated the short_url from the text columns and put it in a new column.
- There are many names in name column called 'None', this word programmatically doesn't Nan so we convert it into Nan, and there are some messy words, I collected the majority of them and convert them in Nan.
- There exist many columns that I don't need it, for example ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog', 'p1_dog')
- Dog stage exist in four columns , so I collected these columns together in one column.
- P1 column contains upper and lower case, so I converted all values into lower case.
- Id_str column should be converted into tweet_id, because I will merge all three tables in one table.
- I converted timestamp column into datetime type.
- There are mistakes in rating_numerator and rating_denominator because rating_denominator must be 10, rating_numerator should not be a large numbers.
- I converted tweet_id column into object type because I don't want it in arithmetic operation.

After I solved all these problems, I merge all these table together in one table, so we can do analysis and visualizing easy.

Up until this point we're doing an amazing job, because we've got our data ready for analysis and visualization.