**Analysis and Visualization for WeRateDog Twitter**

This project is about data wrangle, analysis and visualization. We were given two datasets, a csv file called twitter\_achive\_enhanced, a tsv file called image\_predictions.tsv. The third file, we need to query by ourselves from Twitter API. After we imported all these datasets to pandas data frame, we need to clean the data. Compared to tidiness issue, there are more quality issue for the datasets. Below is the issue I found for the three datasets.

**Quality**

***twitter\_archive table***

* Erroneous datatype: **tweet\_id**, **in\_reply\_to\_status\_id**, **in\_reply\_to\_user\_id**, **retweeted\_status\_id**, **retweeted\_status\_user\_id** should change to string data type, **timestamp**, **retweeted\_status\_timestamp** should change to datetime data type
* Big amount of missing records for **in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id,retweeted\_status\_timestamp**, and small amount of missing records for **expanded\_urls**.
* Rating denominator should always be 10
* Some of the name is "a","an" and "the"
* Rating denominator and rating numerator can be combined in one column called rating.
* Drop **source** and **expanded\_urls** column since they are not useful for the analysis

***image\_predictions table***

* **Tweet\_id** is integer type, should be string type
* Drop **jpg\_url** column

***tweet\_att table***

* Erroneous datatype: **tweet\_id** should be string data type.

**Tidiness**

***twitter\_archive table***

* doggo, pupper, puppo and floofer can be combined into one column called dog\_stage
* Combine rating\_denominator and rating\_numerator together
* Expanded\_url sometimes has two urls in one record

After the cleaning, I combined the three dataset to one master dataset. From there, I can do my data analysis and visualization. I would like to do the analysis by answering some questions I am interested. By answering the question, we know that the stage pupper has lower rating than the other three stages. The confidence level for p1 is much higher than p2 and p3 prediction. For each prediction, the confidence level for True is higher than False.

From the visualization, we can see that the favorite count and retweet count have positive relationship. The more favorite the more retweet.

A screenshot of a cell phone

Description automatically generated