**Wrangle Report**

1. **Introduction**

This project is about gathering data from different 3 sources related to twitter account named WeRateDogs , to wrangle and analyze data about dogs.

1. **Gather Data**

**I gather data using three different resource .**

* At the first dataset I download it from the Udacity , twitter-archive-enhanced.csv , then import it using panda library.
* At the second dataset I also download it from Udacity , ( image-predictions.tsv ) I read it as csv file using panda library.
* At the third dataset I also download it from Udacity , then convert it from Jason file to csv file and read it using json library , with select the most important columns to my analysis ( tweet\_id , favorite\_count , retweet\_count )

1. **Assess Data**

In this part , I explore each dataset , to find the issues in each dataset then try to fix it and finalize it to a clean dataset .

**Assessment Issues**

**Quality**[**¶**](https://viewf6b31853.udacity-student-workspaces.com/notebooks/wrangle_act.ipynb#Quality)

* for df\_1 :
* Names have "a" , "an" , "o" which is not a name.
* The timestamp column is an object. It has to be a datetime object.
* contains retweets and which is duplicates for the data.
* wrong value at rating\_numerator and rating\_denominator.
* missing some expanded\_urls.
* for df\_2 :
* inconsistent capitalization in p1 , p2 , p3.
* There are more than 200 missing value since it has 2075 values insted of 2356.
* for df\_3 :
* There are 2 missing value since it has 2354 values insted of 2356.

**Tidiness**

* one variable in four columns (doggo, floofer, pupper, and puppo) it should be in one column.
* there is a dirty data that should be cleaned in " doggo, floofer, pupper, and puppo "
* There are many columns not needed such as " retweet " related columns.

**Conclusion**

First I gather data , then assess it and last I cleaned it , after I have clean it I store the dataset as new one , so I can visualize it and analyze it easily .