1. **Gathering data**

The first step in data wrangling is to gather data required, so in this project we have multiple types of files.

We have csv file which is abbreviated for comma separated values. We have been given this csv file from Udacity his name is twitter-archive-enhanced.csv. second step is to download tsv file which is named for tab separated values, however we will read differently, we will download the file programmatically. The third step is we will collect some data from twitter API using a library called tweepy, but all of that we needed to import some libraries, of course one of the libraries was pandas, numpy in case needed.

I used read_csv to read twitter archive enhanced and requests library to read a tsv file from url given from Udacity, I used instructions listed to download from twitter API, then I used regular expression to extract needed data such as retweet counts and favorite counts, then I created a data frame containing all data necessary,

2. **Assessing Data**

First I assessed the data visually, I found some things needed to be cleaned and organized. Most of the .in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestam, doggo, floofer, pupper, puppo are empty.

Some names are a or None. And in image prediction the columns' names are confusing.

Then I assessed programmatically the data, I used info method for all data frames and describe method on all data frames, I checked duplicates in all data frames, I used value counts method on names column to view values I found some values equal to a letter, I also checked doggo, puppo, pupper, floofer with value counts method.

I found some quality and tidiness.

3. **Cleaning data**

First I made a copy for all data frames, then I started dealing with the problems needed to be cleaned.

Timestamp is object and should be timestamp, tweet_id should be string, p1, p2, p3 should be categorical.

Tweets counts and favorite counts are string, they should be integer. I used pandas to date time, then I tested the code. Some names are just "a", so I used some regular regression to extract the names then I tested the code. p(1,2,3) in image prediction aren't meaningful, so I changed the names with rename method then I tested the code. Now with tidy issues first I wanted to gather puppo, floofer, pupper all in one column with name dog_type, in order to do this, I used loc method, merge method, rename and drop method. Then I merged image prediction and twitter archive enhanced then I tested the code, I merged favorite count and retweet count with twitter archive enhanced then I tested the code. I split the retweeted status id and user id and time stamp in other data frame. Then split all urls into another data frame.