# Introduction:

This is a wrangle report for the ‘we rate dogs’ project in Udacity’s Data Analysis Professional Track it was done on 3 main steps

# 1.Gathering:

Data was gathered Via 3 sources

1. A CSV file downloaded from udacity given by the data analysis instructors
2. A TSV file for image prediction download programaticaly downloaded from Udacity’s server using a request python package
3. Since i couldn’t API access from twitter therefore i couldn’t gather data from their database using the tweepy library i manually downloaded the JSON file and extracted it using zipfile package in the notebook

# 2. Assessing:

After creating a dataframe for each file, i began to explore each to find Potential issues like tidiness issues or quality issues, i used visual assessing and programmatically assessing some issues were found like:

**Quality Issues:**

- tweet id should be a string
- some denominator values are not equal 10
- timestamp should be coverted to datetime data type
- delete retweets and replys
- drop unneeded columns
- The values in the columns p1_conf, p2_conf and p3_conf should be percentages
- the name column contains None when it should be NULLs
- Unusual names in the name column
- P1, P2, P3 should be categorical

### Tidiness Issues

- doggo, floofer, pupper and puppo columns should be merged into one column
- All dataframes should be in a single dataframe

# 3. Cleaning:

In this step i tried to solve each problem but at first i made a copy of all dataframes to work on to avoid any data loss then

- I changed the type of ‘tweet_id’ in each dataframe to string
- I changed all tweets with denominator not equal 10 to be 10
- I changed time stamp type to datetime
- I deleted rows where there’s an id for reply or retweet
- I deleted the following columns:
    - in_reply_to_status_id
    - In_reply_to_user_id
    - Retweeted_status_id
    - Retweeted_status_user_id
    - retweeted_status_timestamp
- Changed all the None values to be a Null object using np.nan
- Removed all unusual names
- Changed all the values of p1_conf, p2_conf and p3_conf to percentages
- Changed p1,p2,p3 type to category
- Merged the doggo, floofer, pupper and puppo columns into one column named ‘Stage’ then deleted these columns
- Finally merged the 3 dataframes using inner join on tweet_id since it’s the common column in all 3 dataframes into a new CSV file
- I visualised the data to gain some insight

# Conclusion:

The project has been a great experience to learn data analysis and cleaning and even though some problems were discover and solved there are far more advanced issues