# Wrangle Act

First,

## Gathering the Data

Through,

1. **CSV file**: Manual downloading of "twitter-archive-enhanced.csv" then reading it as a Data Frame
2. **Programmatic downloading a TSV file**: downloading "image-predictions.tsv" file then reading it as a Data Frame
3. **API**: Using Twitter API to access the tweet archive of @dos_rates, downloading the data to a txt file the reading it as a Data Frame.

Then,

## Assessing the Data

Programmatic Data Assessment, then documenting the notes found to be cleaned.

Assessment is done separately for each of the gathered data

- "Tweets Archive" Assessment Notes

### Quality Assessment

1. **Wrong Data Type**: Timestamp Column should be converted to a datetime format rather than an object
2. **Unnecessary Columns**: There are un wanted columns for the retweets and text data (5 cols)
3. **Unnecessary Rows**: There are un wanted Rows for the retweets data (181 rows)
4. **Missing Values**: There are missing values in "name" column
5. **Wrong Values**: There are wrong entries in "name" column

### Tidiness Assessment

1. Tweets Data need to be in a separate table
2. Dogs name and ratings need to be in a separate table

- "Images Predictions" Assessment Notes

### Quality Assessment

1. **There are Upper and Lower Case Letters**: The Letters Case in each of the columns "P1, P2, P3" should be the same for the whole column.
2. **Duplicates**: There are tweets with images of similar URLs

### Tidiness Assessment

No Issues

- "Tweets" Assessment Notes

### Quality Assessment

No Issues

### Tidiness Assessment

1. Tweets table need to be merged with other columns related to the tweets in a new table
2. The number of the rows is less than the number of the rows in the other tables to be merged with

Finally,

# Cleaning the Data

That is, solving the notes taken at the Assessment Stage,

## Cleaning Points -- Summary of the Notes Taken at the Assessment Stage

## Quality Issues

**Tweets Archive**

- Wrong Data Type - Cleaning: Convert Timestamp Column data type to be a datetime format rather than an object
- Unnecessary Rows - Cleaning: Drop the unwanted Rows for the retweets data (181 rows)

- Unnecessary Columns - Cleaning: Drop the un wanted columns for the retweets and text data (5 cols)
- Missing Values - Cleaning: Replaced missing values in "name" column
- Wrong Values - Cleaning: Drop the wrong entries in "name" column

### Images Predictions

- There are Upper and Lower Case Letters - Cleaning: Convert all the Letters Case in each of the columns "P1, P2, P3" to be in the lowercase.
- Duplicates - Cleaning: Drop the tweets with images of similar URLs

### Tweets

No Issues

## Tidiness Issues

### Tweets Archive

- Merge dogs' stages in a single column named "Stages"
- Merge tweets-related-data in a separate table
- Merge dogs name and ratings in a separate table

### Images Predictions

- No Issues

### Tweets

- Merge tweets table with other columns related to the tweets in a new table