# DAND_C9

# WeRateDogs

# Wrangle_Report

# Data Wrangling Project


**MOSHIRA EZZAT**
**moshiraezzat@pg.cu.edu.eg**

## Introduction

This is a data wrangling analysis for WeRateDogs Twitter Archive, which is a Twitter account for rating the dog's cool images with humorous comments. These ratings almost have denominator as 10, and numerators that are greater than 10.

## Wrangling Process

### i. Gathering

In this data analysis, the data was gathered as three datasets;

1- The WeRateDogs Twitter archive, manually downloaded as csv file.
2- The tweet image predictions, programmatically downloaded as tsv file.
3- Tweets JSON data queried from Twitter API using tweepy

The three datasets were used to create 3 dataframes by pandas;

1- archive_df,
2- image_prediction_df,
3- api_df

### ii. Assessment

The dataframes were visually and programmatically assessed, to identify quality and tidiness issues summarized in the following table:

| Tidiness Issues | Quality Issues |
|---|---|
| 1- Dog stage represented in 4 columns | 1- Erroneous dtype of tweet_id columns |
| 2- Three predictions and confident columns | 2- Erroneous dtype of timestamp |
| 3- Inconsistent name of id column in api_df | 3- Presence of retweets |
| 4- Data separated in three tables | 4- Tweets without images |
| | 5- Unnecessary columns |
| | 6- Invalid values in name column |
| | 7- Tweet's sources are masked |
| | 8- Ratings are not calculated |

### iii. Cleaning

First observation was the presence of inconsistent name of *'id'* column in api_df, that is renamed into *'tweet_id'* to be consistent with that of archive and image-prediction dataframes.

The dtypes of '*tweet_id'* columns were 'integr' in all dataframes, as well as the dtype of *'timestamp'* column in the archive_df was 'string'. These issues were solved using astype() function to convert tweet_id from integr to string, and to_datetime() function to convert the timestamp dtype from string to date time.

Contrarious to the tidiness rules, the dog stage in archive_df was represented in four separated column, *['doggo','pupper', 'floofer', 'puppo' ],* so the four columns were combined into one column called dog_stage.

Similarly, the columns of predictions *p1,p2,p3* and its confidence *p1_conf, p2_conf, p3_conf* in the image_prediction_df were unified into single prediction column according to the highest confidence and the boolean values in the *p1_dog, p2_dog, p3_dog* columns.

One of the principal key points for this project is to analyze original tweets, not the retweets or replies. Through the assessment step, it was obvious the presence of non-null records in the *in_reply_to_status_id*, and *retweeted_status_id* columns of the archive_df; referring to the presence of retweet and replies. So, these non-null records were dropped, then all retweets and replies columns *['in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp']* were dropped totally since they are unnecessary in this analysis.

Also, tweets missing *expanded_urls* in the archive_df, were dropped as they were actually missing the image itself.

The invalid values in the name column of archive_df were replaced with 'None'

The sources of the tweets were masked at the end of html string. The clean types of sources were extracted and displayed clearly.

The rating system of WeRateDogs follows a scale of 1 to 10, but it is invariably may have ratings exceeds maximum, so it was decided to only calculate the ratings by dividing the numerator by 10 in a new column *[ratings]*. Then the *rating_numerator* and *rating_denominator* columns were dropped.

At the end of the cleaning steps, the dataframes were merged into a master_df using *tweet_id* and storied as *twitter_archive_master.csv* for further iteration of the wrangling process and data analysis.