# Reporting: Wrangle Report

## Data Gathering

- I imported all the necessary python libraries (pandas, numpy, matplotlib, seaborn, tweepy, requests, json, timeit) with their aliases into the notebook.
- I required 3 datasets to complete this analysis but only one was readily available in csv format so I read the dataset available which was twitter-archive-enhanced.csv into a dataframe using pandas.read_csv method.
- The second dataset, image_prediction.tsv was downloaded programmatically with the requests library using the url that was given and then read into a dataframe using pandas.read_csv method.
- The third dataset was scraped from twitter using twitter API and a python module named tweepy. This was used to query twitter's database for information on tweets from WeRateDogs twitter page dating to August 2017.
- This data was extracted into a json file and was read into a dataframe using the pandas.read_json method. I then extracted the necessary columns in the dataframe by subsetting it for columns that had information on tweet_id, retweet count and favorite count.

## Data Assessment

- Data from the three datasets were assessed programmatically using various methods such as df.head, df.info, df.describe, df.isnull, df.duplicated, df.value_counts, df.query, df.quantile, df.unique, df.nunique to check for tidiness and quality issues in the dataframe.
- After each assessment, I documented the quality and tidiness issue I found and then moved on to the cleaning stage.

## Data Cleaning

- The cleaning stage is divided into two parts: Quality issues cleaning and Tidiness issues cleaning.
- Before I started cleaning the dataframes, I made a copy of each dataframe in order to preserve the original state of the dataset.
- For each quality issue documented to be cleaned I divided the cleaning into three stages: Define(where I defined the method of cleaning), Code(the cell which was used to actually make positive changes into the dataframe to make it better for analysis) and Test(where i checked if the changes I made to the dataframe worked). I then went ahead to use this same iterative method for the tidiness issues.
- When I completed the extensive cleaning of the three dataframes, I combined the clean dataframes into one dataframe and named it twitter_archive_master and stored it in twitter_archive_master.csv file.