# PROJECT: WERATEDOGS DATA WRANGLING AND ANALYSIS REPORT

## INTRODUCTION

I conducted a data analysis project focused on dog ratings on Twitter. I worked with three different data sources, including data obtained via the Twitter API (gotten from udacity), and used Python and the pandas library to carry out data wrangling tasks, including merging prediction columns, changing data types, removing retweets, and addressing inaccurate ratings. I also cleansed the data to improve its accuracy and readability, and capitalized dog breed names that were not truly dogs. Additionally, I used the matplotlib and seaborn libraries to visualize the data, gaining insights such as the most liked and retweeted dog breeds based on favourite and retweet counts, the distribution of dog breeds, the distribution of tweet sources, and the month with the highest frequency of tweets.

### OBSERVATIONS MADE FROM THE FIRST DATASET (twitter-archive-enhanced.csv)

- The timestamp column is a string type and should be in datetime data type
- The Dog stages should be in one column
- The source of the tweet is ambiguous and not easy to understand
- Numerator and Denominator rating values are improperly
- Only original tweets with pictures are required but some of the tweets are retweets
- Some Columns will be dropped as they have alot of missing values

### OBSERVATIONS MEDE FROM THE SECOND DATASET (image-predictions.tsv)

- Duplicated jpg urls
- Some dog names need to be capitalized and some dog names are not real names
- prediction columns can be merged into one standard prediction Rating

QUALITY CONCERNS

- Merge data prediction columns to create a single column with dog breed and confidence ratings for predictions
- The timestamp column should be changed to a `datetime` data-type.
- Retweets will be deleted, we need original tweets with images.
- The provided source data is challenging to read and has to be cleansed
- Inaccurate ratings in the denominators and numerators
- Since several of the columns' values are lacking, some of them will be removed.
- Use to fill in the blanks in the expanded urls column. `dropna`
- Make all necessary corrections to the dataset's incorrect data types, including switching the source and dog stages data types to the category data type for later analysis.
- Dog breeds that aren't truly dogs should have their names capitalized.

TIDINESS CONCERNS

- The `dog_stages` column for `df1` `twitter_archive_enhanced.csv`
- The information about the tweets is dispersed across three separate datasets and can be fit into a single column in csv.

# CONCLUSION

My goal was to improve the accuracy and usefulness of the data for further analysis, and I believe that my work achieved this goal while also providing valuable insights into dog ratings on Twitter