# Wrangling Report

This is the data wrangling report, it summarizes the steps and efforts taken during the data wrangling project.
The dataset of Twitter user @dog_rates, also known as We Rate Dogs was wrangled ( and analyzed and visualized). The account rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, however, are almost always greater than 10. WeRateDogs has over 4 million followers and has received international media coverage.
This entire project was completed on my local device, however, the wrangle_report and act_report were completed on Google docs.

These were the Three(3) steps taken;

Step 1: Gathering data
Step 2: Assessing data
Step 3: Cleaning data
Step 4: Storing data
Step 5: Analyzing, and visualizing data
Step 6: Reporting

*This is a breakdown of all the steps that were taken in order to ensure accurate data wrangling process;*

## Step 1: Gathering Data

#### The WeRateDogs Twitter archive

This was the first step in this step.I downloaded 'The WeRateDogs Twitter archive'file manually by clicking the following link: `twitter_archive_enhanced.csv`. Once it was downloaded, I uploaded it to my local device and read the data into a pandas DataFrame.

**The tweet image predictions**
This was obtained by running every image in the We Rate Dogs Twitter Archive through a neural network that can classify dog breeds. It resulted in a table that was full with the top 3 image predictions alongside tweet ids, image urls and image number that corresponded to the most confident prediction.
This file (`image_predictions.tsv`) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

**Additional data from the Twitter API**
This was obtained by querying the Twitter API and it was then stored in a txt file called tweet_json with my twitter development account.

The next step was assessing the data;

**Step 2: Assessing data**
After the gathering process was completed, I began assessing the data both visually and programmatically for both quality and tidiness issues.

The following were the concluded findings;

# a. Observations For Enhanced Twitter Archive

## Tidiness

* Dog names are not consistent

## Quality
* There are retweets present in the data

* ID variables are sometimes integers or floats (numeric)
* Column names are not always meaningful
* Retweeted_status_timestamp is not a datetime variable
* Source values are formatted as <a href=url <a/>
* Some rating numerators less than 10
* "retweeted_status" variables are numeric

# b. Observations For Tweet Image Predictions

## Quality
* Some column names contain '_' and '-' instead of spaces
* Some names start with an uppercase while some start with a lowercase

# c. Observations For Twitter API

### Quality

- Missing data (the archive dataset has 2356 ids but only 2354 show up)

Note: Not all of these issues were cleaned

**Step 3: Cleaning data**

# Cleaning Data

The issues stated above were cleaned appropriately resulting in a tidy data pandas DataFrame.

**Step 4: Storing data**

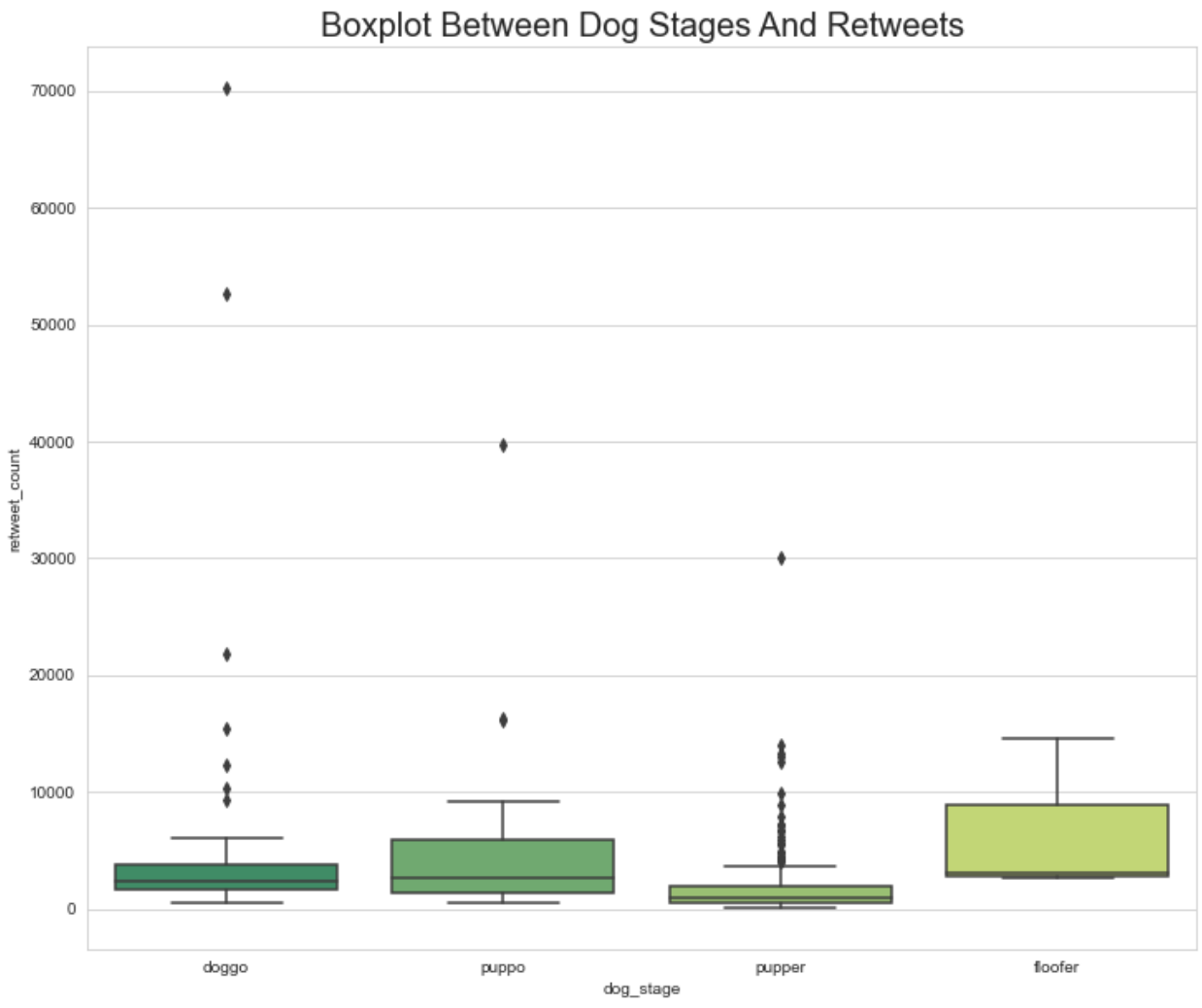
I stored the cleaned master DataFrame in a CSV file with the main one named `twitter_archive_master.csv.`

# Step 5: Analyzing, and visualizing data

## I. Relationship Between Favourite_count and Retweet_count

## Analysis

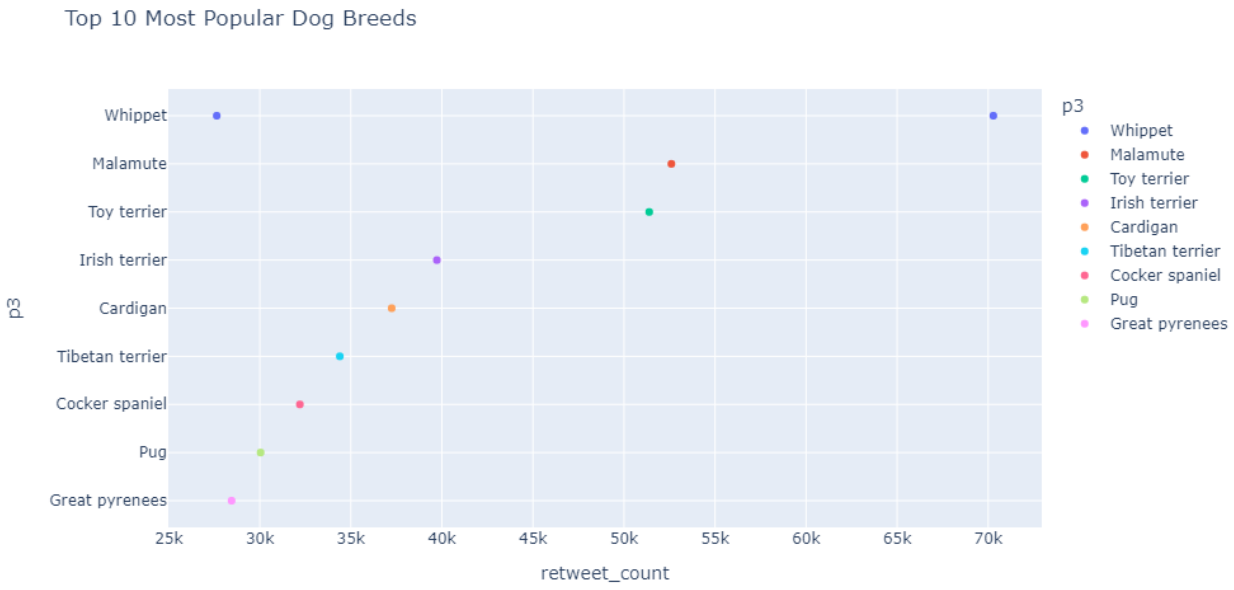
- There are more dogs in puppo category
- There are more retweets in doggo category



## II .Top 10 Most Popular Dog Breeds

## Analysis

- 'Whippet is the most retweeted dog breed



## III . Amount of Each Dog Stage

## Analysis

- Pupper has the highest percentage
- Floofer has the least percentage