Attempting this project was a great challenge as I learned how to solve problems better and understood the wrangling process. The following are the step I took to **gather**, **assess**, **clean**, **store**, **analyze** and **visualize** the data:

Gathering:

I gathered 3 sets of data with 3 different file types; csv, tsv and json.

- Downloaded the 'twitter_archived_enhanced.csv' file which was provided by We Rate Dogs through Udacity.

- Getting the 'image_predictions.tsv' file was relatively easy as I used programmatic method to download, rename and save it from the site provided.

- For the twitter data, I had to apply for an API access which helped me access 'favorites' count and 'retweets' count tweets for specified tweet IDs. It was converted into a dataframe

Assessing:

I assessed the datasets visually and programmatically. I discovered some quality and tidiness issues which I later cleaned:

Quality

archived table

- NaN values present in different columns ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp').
- Names; Identified non-typical names like 'a', 'such', quite', 'the'.
- Columns such as; in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp have incomplete rows when compared to the total number of rows as as 'expanded_url' column.
- rating_denominator: Identified rating denominators less than and greater than 10.
- timestamp: It would look cleaner if we removed the timezone value.
- 'None' is used to represent a null value.

image_predictions table

- Irregular name of breeds; some are in lower case, while other in upper case, others have an underscore between 1st and second names.
- Column name not descriptive

tweets table

- No quality issues

Tidiness

- The four dog stages have different columns in the archived table which can be represented in just 1 column (dog_stage).
- Some columns in table are not needed for analysis which distracts the eyes and makes it look untidy.

Cleaning:

This was a challenging but enlightening part of the project. I used the format specified; **define**, **code** and **test**. All issues specified above were cleaned. 'twitter_archived_enhanced.csv' and data scrapped from twitter were merged into one.

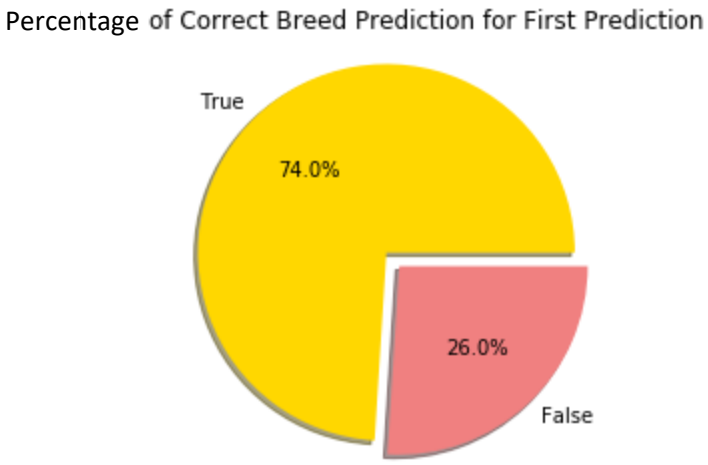
Storing:

After cleaning, the dataframes were stored as .csv files.

Analyzing:

Looking at all the dataframes, I was curious to know the accuracy of the algorithm used to predict the dog breeds 'p1_dog', 'p2_dog' and 'p3_dog'. I programmatically found out the different percentage accuracy of the 1st, 2nd and 3rd predictions. All predictions had above 70% 'True' values and between 20 – 30% 'False' values with the 3rd prediction having a larger share of the 'False' values. For the merged dataframe, I wanted to the names of the dog with the 10 highest favorites count. Most of the names where unknown.

Visualizing:



Since this made me curious, I decided to create a visualization showing clearly the percentage of true to false predictions.

Conclusion:

This has been undoubtedly the most exciting project I've done with Python. I will look further into the data and my code to create better versions of it. I feel like I can handle data wrangle now but will improve with time.