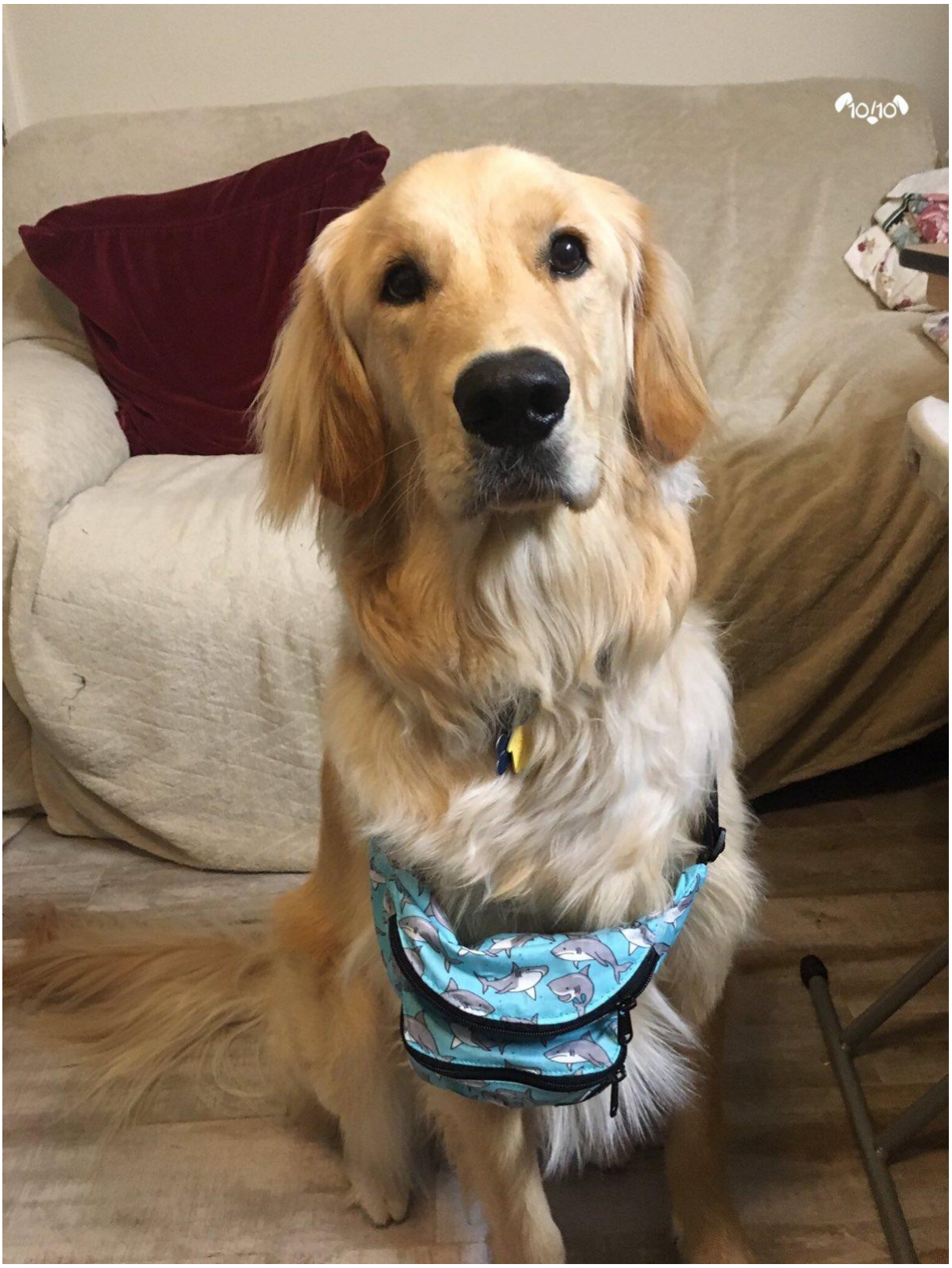
WE RATE DOGS DATA WRANGLE REPORT WITH SOME DOGS PICTURES FROM TWEETS



This is Stuart. He's sporting his favourite fanny pack secretly filled with bones only

The following steps were involved in the Data Wrangling and analyzation project:

Gathering of Data:

I downloaded the twitter-archive-enhanced.csv file from the link provided in my classroom.

I created a folder on my desktop named image_predictions and I downloaded the image_predictions.tsv file to it.

I downloaded tweet_json data and the twitter_api.py file from the link provided in my classroom .

I uploaded all the data into my jupyter notebook.and imported my libraries.

Assessing of Data:

This was done using both visual and pragmatic assessment.

I read the gathered data into various data frames.

I am meant to detect and document at least 8 quality issues and 2 tidiness issues as indicated below which has been broken down to their various Data Frames:

Quality:

1. **Twitter Archive Dataset:**

- There are 181 retweets duplicated in retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
- There are 78 reply tweets in_reply_to_status_id, and in_reply_to_user_id.
- The column rating_numerator had some values wrongly extracted
- There are 2297 tweets with expanded_urls.
- Timestamp is in string format

- Name column have invalid names i.e. None and a
- Several other columns have NaN and none values
- There are 745 tweets with the dog name as "None" and there are also other names that are just alphabets with no meaning

2. **Image Prediction Dataset**
   - The file downloaded contained some joined strings
   - There are 2075 image predictions, 281 less than the number of tweets in the Twitter archive dataset.

3. **Tweet Json Data:**
   - Missing values found when reading Json Data., leaving me with 2327 tweets which is 29 less than the number of tweets in the Twitter archive dataset
   - High and out of range values for numerator and denominator rating, with maximum for both being 1776 and 170 respectively

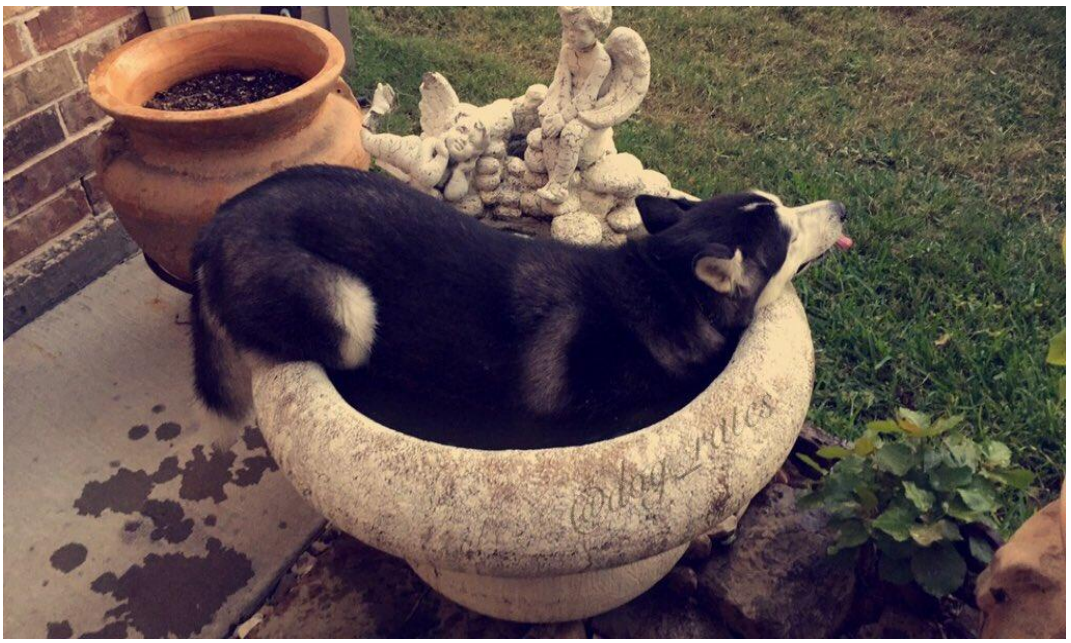
Tidiness:

- We are interested in only the original tweets for this project,so the columns retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp could be done without.
- Twitter Archive Data - doggo, fluffer, pupper and puppo column all pointing to dog stages were in separate columns..
- Also a new column was created for dog_gender
- Image Prediction - The columns p1, p2 and p3 contain the same type of data, predictions.
- Tweet Json Data - rating denominator column was deleted when all denominator was cleaned to the same value 10.
- Some column names were repetitive

Cleaning of Data:

Cleaning of data was done both manually and programmatically.
The following were some of the steps taken to clean my data:

- Duplications were found and deleted and I also filtered out the retweeted_status_user_id that had NaN and no images.
- By dropping columns not needed for my analyses
- By correcting some erroneous data types and and also converting columns to matching data types so that analyses can be done.
- Missing values found in name and generated column dog_stages were converted to none
- Name had values that were not correct this were extracted and changed to none
- Correction of the values of the column rating_numerator that had some values that were wrongly extracted
- Dropping of the column with numerator_rating of 0
- Tweets text had some string characters that were removed
- some joined strings that were separated using the sep function and regex
- Tweet Json Data - rating denominator of value 10 was adopted for the analysis
- Tweet Json Data - rating denominator column was deleted as it has the same value 10.
- Some columns (timestamp': 'tweet_date', 'source': 'tweet_source', 'text': 'tweet_text', 'expanded_urls': 'tweet_url', 'jpg_url': 'tweet_picture', favorite_count': 'tweet_favorite', 'retweet_count': 'tweet_retweet',
  'prediction_algorithm' : 'dog_breed') were renamed as indicated to make it look tidier.



This is Jimison. He's stuck in a pot.