Data Wrangling

Objectives:*(The traffic light from data perspective)*

1. **Gathering** 2. **Assessing** 3. **Cleaning**

Step 1: Gathering

The first step in the process is gathering the data which serves the analysis. In this project, there is 3 different source of data gathered. And are as the following:

1. The **twitter archive** which is in **csv format** and downloadable from udacity as it was sent by the twitter team. Therefore, it contains the data about **@WeRateDogs**.
2. The **image prediction** neural network data outcome in a **tsv file** which can be downloaded by the request library or from udacity.
3. The **twitter api** data which can be gathered by the tweepy lib. but there must be a developer account on twitter api. Also it is downloadable from udacity as a **txt file**.

Step 2: Assessing

Assessing the data is the process of observing the data programatically using python and visually using g-sheets. The target is to discover the quality and tidiness issues. The following problems are the outcome of the assessing and reassessing process the data has been observed and reobserved multiple times.

Quality Issues

- Dataset 1 Dropping the rows with values in retweeted_status_id before dropping the column (UPDATE)
- The timestamp column should be datetime format.
- The dog (name,doggo,floofer,pupper,puppo) columns have Nones instead of NaNs.
- The source html format in source column since the source can be extracted.
- "A Retweet is a re-posting of a Tweet.". So its columns can be dropped since it's not the interest.
- In reply is also to be dropped.
- The name column in 1747 index is 'officially' which doesn't match.
- Some of the numerator ratings inside text columns are in decimal @index 45.
- The text column is including the rating and the review with respect to the link.
- Sometimes denominator ratings are not out of 10.
- The NaNs in expanded urls. Since expanded urls = tweet URLs which is essential.
- Some dogs are classified into multiple stage.
- Drop the tweets that are NaNs.
- Remove all of the '.*only rate dogs' of the tweet column as shown in index 25.
- Dataset 2
- Drop the img_num as for example each img_num category have different links and are showing different results.
- Combining p1,p2,p3 and p1_conf,p2_conf,p3_conf ,p1_dog,p2_dog,p3_dog into 2 columns only (dog name & probability).Drop the unwanted columns p1,....p3_dog as we already got the dog_name and dog_prob.
- Drop the NaN if all of the p (p1,p2,p3) that are false since the false combined indicating that this row does not belong to any dog.

- Dataset 3
- Problem in naming the id column should match the other 2 dataset should be tweet_id instead of id.
- Any column except tweet_id, favourite_count, retweeted_count can be dropped.

Tidiness Issues

- Dataset 1
- (Already combined) columns (doggo,floofer,pupper,puppo) to dog_stat .Drop the unwanted columns (doggo,floofer,pupper,puppo).
- Combining the data based on tweet_id.
- Combining all of the data into 1 dataset.

Step 3: Cleaning ***The last step in the process is cleaning and visualize it based on step 2 which was based on step 1. The amount of cleanliness will show better results in visualization. The cleaning steps are as following:*** ****

1. **Define the problem based on each observation.**
2. **Code programatically.**
3. **Test the final product.**