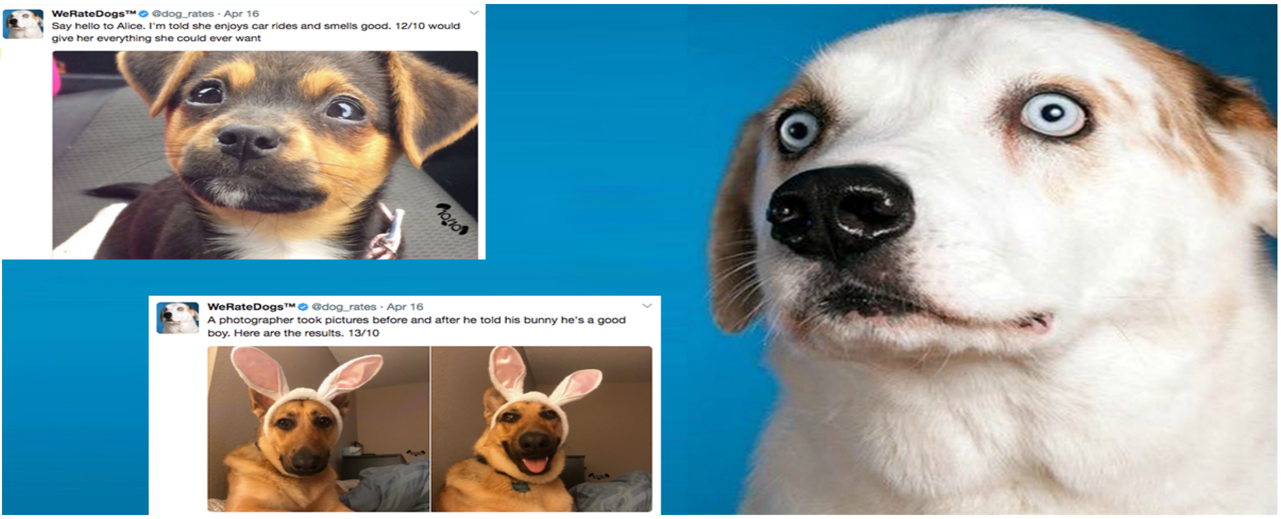
WeRateDogs

user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.



WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Wrangling Report

Gathering part

Data is successfully gathered:

- From at least three (3) different sources .
- In at least three (3) different file formats on the Project .
- Each piece of data is imported into a separate pandas DataFrame at first.

Obtaining data

- Getting data from an existing file (twitter-archive-enhanced.csv) Reading from csv file using pandas.
- Downloading a file from the internet (image-predictions.tsv) Downloading file using requests.
- Querying an API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy.
- Importing that data into our programming environment (Jupyter Notebook).

Assesing part

- **Visual assessment**: scrolled through the data using my preferred software application (Excel).
- **Programmatic assessment:** used code to view specific portions and summaries of the data (pandas' head, tail, and info methods).

Assessed data for by checking both quality and tidiness :

- **Quality:** issues with content which indicate low quality data.

-Main problems that I have tackled was the deviation and high rating that was represented in the arc df, i had to check the validity of them by sorting the illogical values also there was alot of missing values.

-The presented screenshot ahead summaries the quality issues found on each data used.

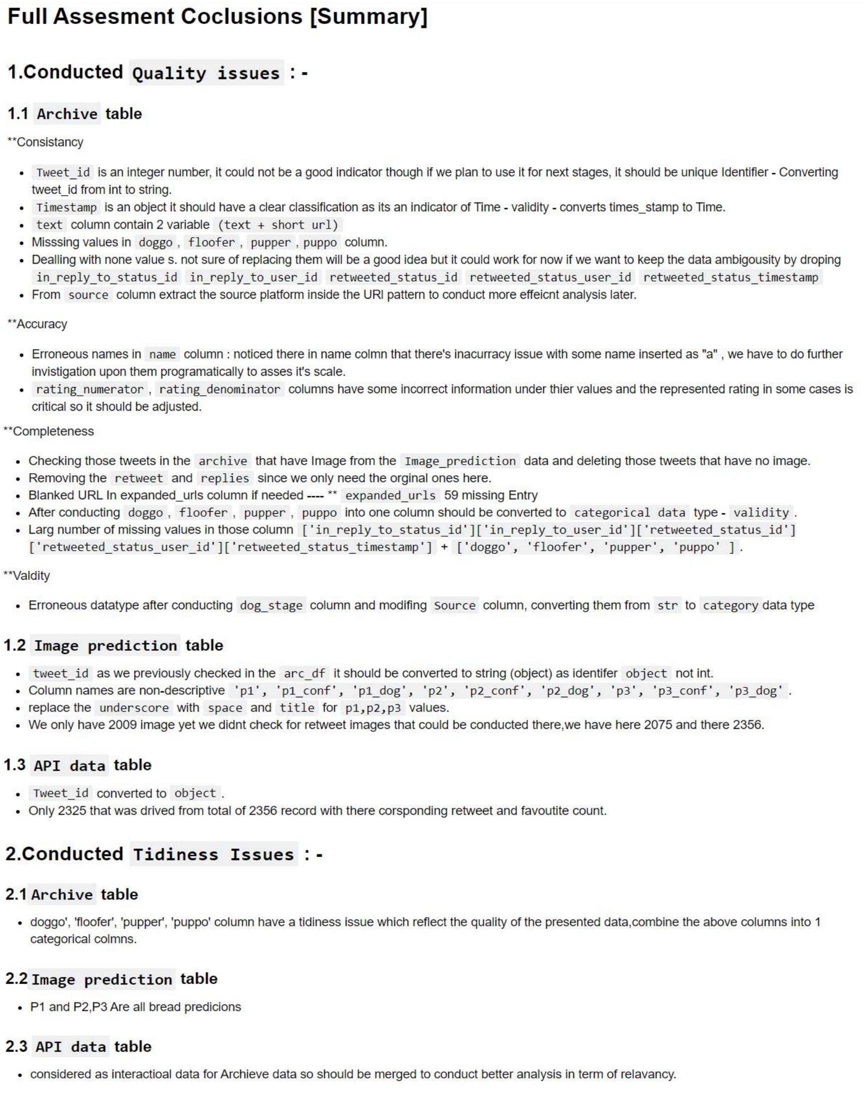
- Tidiness : issues with structure that prevent easy analysis.

- I followed the tidy data requirements for detection as :

- ✓ Each variable forms a column.
- ✓ Each observation forms a row.
- ✓ Each type of observational unit forms a table.

so i had to ensure that they are meeting those creteria and in good shape to be represented.

- Screenshot of the conducted assesment summary



Full Assesment Coclusions [Summary]

1.Conducted `Quality issues` : -

1.1 `Archive` table

**Consistancy

- `Tweet_id` is an integer number, it could not be a good indicator though if we plan to use it for next stages, it should be unique Identifier - Converting tweet_id from int to string.
- `Timestamp` is an object it should have a clear classification as its an indicator of Time - validity - converts times_stamp to Time.
- `text` column contain 2 variable `(text + short url)`
- Misssing values in `doggo` , `floofer` , `pupper` , `puppo` column.
- Dealling with none value s. not sure of replacing them will be a good idea but it could work for now if we want to keep the data ambigousity by droping `in_reply_to_status_id` `in_reply_to_user_id` `retweeted_status_id` `retweeted_status_user_id` `retweeted_status_timestamp`
- From `source` column extract the source platform inside the URl pattern to conduct more effeicnt analysis later.

**Accuracy

- Erroneous names in `name` column : noticed there in name colmn that there's inacurracy issue with some name inserted as "a" , we have to do further invistigation upon them programatically to asses it's scale.
- `rating_numerator` , `rating_denominator` columns have some incorrect information under thier values and the represented rating in some cases is critical so it should be adjusted.

**Completeness

- Checking those tweets in the `archive` that have Image from the `Image_prediction` data and deleting those tweets that have no image.
- Removing the `retweet` and `replies` since we only need the orginal ones here.
- Blanked URL In expanded_urls column if needed ---- ** `expanded_urls` 59 missing Entry
- After conducting `doggo` , `floofer` , `pupper` , `puppo` into one column should be converted to `categorical data` type - `validity` .
- Larg number of missing values in those column `['in_reply_to_status_id']['in_reply_to_user_id']['retweeted_status_id']['retweeted_status_user_id']['retweeted_status_timestamp'] + ['doggo', 'floofer', 'pupper', 'puppo' ]` .

**Valdity

- Erroneous datatype after conducting `dog_stage` column and modifing `Source` column, converting them from `str` to `category` data type

1.2 `Image prediction` table

- `tweet_id` as we previously checked in the `arc_df` it should be converted to string (object) as identifer `object` not int.
- Column names are non-descriptive `'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'` .
- replace the `underscore` with `space` and `title` for `p1,p2,p3` values.
- We only have 2009 image yet we didnt check for retweet images that could be conducted there,we have here 2075 and there 2356.

1.3 `API data` table

- `Tweet_id` converted to `object` .
- Only 2325 that was drived from total of 2356 record with there corsponding retweet and favoutite count.

2.Conducted `Tidiness Issues` : -

2.1 `Archive` table

- doggo', 'floofer', 'pupper', 'puppo' column have a tidiness issue which reflect the quality of the presented data,combine the above columns into 1 categorical colmns.

2.2 `Image prediction` table

- P1 and P2,P3 Are all bread predicions

2.3 `API data` table

- considered as interactioal data for Archieve data so should be merged to conduct better analysis in term of relavancy.

Cleaning part

I used various types of technique to clean the data appropriately using python and it's related packages through jupyter notebook to ensure that the data will be cleaned and ready to be used later to conduct more insightful conclusions about it.

Analysis and Visualization part

at the end I Communicated the insights and displayed the visualizations produced from my wrangled cleaned data.