Wrangle report

The dataset that wrangles, analyzed and visualized is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog." WeRateDogs has over 4 million followers and has received international media coverage.

The wrangling process took place in main three steps:

1. Gather data
2. Assess data
3. Clean data

The gather steps

It was conducted to collect the data from

- Twitter API by a developer account and python teetpy library

- the WeRateDogs Twitter archive, which was given by Udacity as a on-hand file.

- Dog’s images predictions, i.e., what breed of dog (or other object, animal, etc.) was present in each tweet according to a neural network. This file was hosted on Udacity's servers and should be downloaded programmatically using the Requests library.

The Assess steps

It was conducted manually and programmatically, and the quality and tidiness issues found was:

1. Quality issues

Image Prediction data

- There is duplicates in column image url
- Upper case letters in columns p1, p2, p3

Twiter API data

- Unnecessary columns for analysis: id_srt, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, entities, truncated, user
- Null columns: coordinates, place, contributors, is_quote_status
- created at should be timestamp
- id should be tweet id for coherency
- 'lang' column should be category datatype

Twitter Archive data

- No need for columns of retweet info. and text and replies
- timestamp column is date not string
- Doges names missing and wrong like "a", "very"

2. Tidiness issues

- The dogs breed should be in one column

- twitter data in the three dataframes separated

The Clean steps

The cleaning steps was showed in the format of Udacity course as

- **Define** the cleaning act
- **Code** to solve the problem
- **Test** if the code works

In order to clean the previous mentioned issues in the data the below steps were taken:

- Remove all duplicates from image url column
- Transform all the columns instances into lowercase data
- Drop all unnecessry columns
- Remove null columns as it have the same instance
- Change created_at to timestamp for consistency of entities
- Change id column name to tweet id for coherency
- Change lang column to be category datatype
- Remove columns of rewteet info. and text
- Change timestamp column from string datatype tp date
- Remove corrupted names
- make one column with dogs breed only
- Merge the three dataframes