WRANGLE REPORT

INTRODUCTION

- This project based on a Twitter account that rates people's dogs with a humorous comment about the dog named WeRateDogs. 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.
- I used this account to gather data asset and clean to analyze these tweets and visualize it.

GATHERING

- There was three resource to gather data
- Twitter archive :

twiiter archive was a csv file privided by udacity ,I use pandas to read it by this function

pd.read_csv('**twitter_archive_enhanced.csv**').

GATHERING

- Tweet Image prediction :

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and downloaded programmatically using the **Requests** library

GATHERING

- Twitter's API
- It was suppose to me to query the Twitter API for each tweet's JSON data using Python's **Tweepy** library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data written to its own line. Then read this .txt file line by line into a pandas Data Frame with tweet ID, retweet count, and favorite count, so I asked for permission from twitter but they rejected my request .so I read the file by pd.read_json('tweet_json.txt').

ASSESSING

- Visual assessment:

By show the whole data frame and sneak a peek to the data and discovered it.

- Programmatic assessment:

pandas' functions and/or methods are used to assess the data. such as :

df.describe()

df.info()

Df.head()\df.tail()

ASSESSING

- Quality issues:

-unused columns such as "retweeted_status_timestamp,in_reply_to_status_id,in_reply_to_user_id,retweeted_status_user_id

-date stamp in str

-tweet id in int

-keep only the original rating(no retweet)

-rating denominator not equal 10

-duplicated data

-underscore in p1,p2 and p3

-rename id to tweet id

- Tidiness issues :

-merge dog stage "doggo,floofer,pupper,puppo"

-merge tables

CLEANING

- After I collect the data and assess it, I clean all decomented issue by define the problem ,write the correct code to clean it then test it.