WRANGLE REPORT
BY ISHAN ARORA

Introduction:

The dataset that we will be wrangling (and analyzing and visualizing) 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. 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.

Wrangling Summary:

Basically wrangling is divided into 3 categories-
Data Gathering
Data Assessing
Data Cleaning

Data Gathering

As the name suggests in this process we have to gather data from different sources. In our case we gathered from three different sources

Twitter Archive File: twitter_archive_enhanced.csv was provided by Udacity so it was a local file
The tweet image predictions: It was hosted on Udacity servers. It consisted of breed of dogs present. It was downloaded programmatically using Requests Library and URL function
Twitter Library and JSON: I used python Tweepy library. I stored the json in text file and then read it converted into a dataframe

Data Assessing

Next up is the assessing part. It is the most fun part for me as here we become detectives and find things that needed to be corrected about the data. I assessed the data as follows:

- I looked over the excel files and printed the dataframe in Jupyter Notebook and tried to find some observations about the panda
- I also used the useful pandas describe and info functions to get more insight about data, check for missing values
- I also checked for duplicated value

At last I wrote down all the quality and tidiness issues that I inferred from my observations

Data Cleaning

This part is further divided into 3 parts-Define, code , test
1)First I created a copy of original dataframe. This was done so that in case of errors the original dataframe isn't effected.
2)Then I handled all the Quality issues.There were many quality issues that were found by me.

- I started with changing of datatypes
- Then removed the retweeted tweets
- Changed the missing values into NAN
- Adjusted Numerator
- Capitalized Breed names

Etc.
3)Then there were tidiness issues like having unnessary columns and at last merging all the dataframes into one calling it the master dataset
4) Lastly I stored my master dataset into external csv file

Conclusion:

Data Wrangling is one of the most important skill that one should be familier with. There were many skills that were covered in this project form gathering to cleaning. Pyhton pandas library is a great tool and help a lot with this process along with numpy and matplotlib