INTRODUCTION

This is a project that involves data wrangling of 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. The data wrangling process involves three key steps, i.e.

1) Gathering data
2) Assessing data and,
3) Cleaning data

My analysis begins with importing various python libraries that will be used to gather, analyse, clean and develop insights from the data. The libraries used are pandas, numpy, requests, tweepy, json, seaborn, matplotlib, plotly and IPython.display.

STEP ONE: GATHERING DATA

The project has 3 datasets, i.e.

- Enhanced Twitter Archive (df1) – Provided by Udacity in csv format and is loaded into the notebook using the *read_csv* function
- Image Predictions File (df2) – Provided by Udacity and is loaded using the requests library.
- Additional Data via the Twitter API (df3).

The Additional Data via the Twitter API required first requesting access to the Twitter API via a developer account that would allow access to token keys required in gathering of the data. I then used the tweet IDs in the Twitter archive dataset to query the Twitter API for each tweet's JSON data and stored each tweet's entire set of JSON data in a file called tweet_json.txt. After which I read the json file into a pandas data frame.

STEP TWO: ASSESSING THE DATA

I first assesses the data visually simply by scrolling through the datasets, I noted that some of the columns and rows had missing values.

To assess the data programmatically, I ran the below codes across the 3 datasets:

- df.shape – to check the size of the dataset in terms of number of rows and columns
- df.info() – to check the datatypes of the columns
- df.describe() – check the descriptive statistics of the data
- df.isnull().sum() – to check the sum of null values per column
- df.duplicated().sum() – to check the sum of duplicated values if any
- df.sample() – to check a random sample of 10 rows of the data

I also assessed various columns individually.

Some of the quality issues that I was looking out for included missing data, incorrect data types, duplicated data, irrelevant columns/rows, etc.

Some of the issues identified were as below:

1. **Quality issues**

Dataset 1:

- Incorrect data type for column 'timestamp' - should be datetime

- Incorrect data type for column 'tweet ids' - should be string as this is an identifier
- Remove rows that have retweet status ID as these are not original ratings
- Irrelevant columns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls'
- Rows with text that contains "We only rate dogs" in the 'text' column are not dogs.
- Values with 'None' instead of 'NaN' in the columns: 'doggo', 'fluffer', 'pupper' and 'puppo'
- Incorrect extraction of some rating_numerator values
- Incorrect data type for column 'rating_numerator' - should be float

Dataset 2:

- Incorrect data type for column 'tweet ids'- should be string as this is an identifier.
- Rows where the number one predictions (i.e. column 'p1_dog' is false) are not dogs.

Dataset 3:

- Incorrect data type for column 'tweet ids'- should be string as this is an identifier
- Inconsistency in column name to identify tweet ids. Change 'id' to tweet_id'
- Irrelevant columns not needed in my analysis

2. **Tidiness issues**

- 1 variable(dog type) in 4 columns i.e. doggo, fluffer, pupper, puppo
- Combine all three datasets

STEP THREE: CLEANING DATA

The first step in this process involved creating copies of the data that is being cleaned, hence I created copies of the 3 datasets and named them *df1_new, df2_new* and *df3_new.* I used the define, code, test method in my cleaning process, where I defined the issues raised above, wrote the code to correct the issue then performed a test to confirm that the issue was indeed corrected. This process has been well documented within the Jupyter Notebook.

STORING DATA

After the cleaning process, I stored my data in csv format in a file named:
'twitter_archive_master.csv'.