WRANGLE AND ANALYZE DATA
DATA ANALYST NANODEGREE PROJECT

INTRODUCTION

The purpose of this project is to wrangle and analyze a dataset. The dataset is a tweets archive from Twitter with user @dog_rates. It is a twitter account that rates people's dogs. This project describes the process of data wrangling, visualizations and provides conclusions based on my findings.
This report contains:

1. Wrangling data by gathering, assessing and cleaning data
2. Storing, analyzing and visualizing the wrangled data
3. Reporting on the data wrangling and data analysis

GATHERING DATA

1. Twitter archive file: twitter-archive-enchaced-2.csv was already provided by Udacity and was downloaded manually.
2. Tweet Image Predictions: image_predictions.tsv file was hosted on Udacity's website and was downloaded using the request library and URL information given. It mentions what breed is present in each tweet (according to a neural network)
3. Twitter API and JSON: By using the tweet IDs. In the twitter archive, I queried the twitter API for each tweet's JSON data using Python's tweepy library and stored each tweet's entire set of JSON data in a file (tweet_json.txt)

ASSESSING DATA

The data obtained was assessed for quality and tidiness issues.
Three data tables were obtained in the gathering information stage.

1. Quality issues:
    - Missing data present in in_reply_to_status_is, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id etc.

- Missing dog names. Some of the dog names are just ‘a’ or ‘an’
- The dataset includes retweets, which indicates duplicate data
- The timestamp was an object.
- Rating_numerator should also be within 10. But its value is more than in many observations
- rating_denominator should be a standard 10 but there are many other values
- The source column has the HTML tags
- p1, p2, p3 columns in images have invalid data. They also contain data separated by underscore for names

2. Tidiness issues:
    - Doggo, floofer, pupper, puppo columns in the twitter_archive_enhanced.csv should be combined into a single column as this is one variable that identify stage of dog
    - The images data is a part of the same observational unit as the data in the twitter archive.
    - Information about one type of observational unit (tweets) is spread across 3 different files/dataframes. So these 3 dataframes should be merged as they are part of the same observational unit.

Visually, the dataset can be assessed for column names, data types etc in the Jupyter notebook or in an excel spreadsheet.
Programmatically the data can be assessed by using methods like .info(), .value_counts(), .duplicated, etc

CLEANING DATA

1. Twitter archive contained dog stages in four different columns which has to be made into one single column.
2. Deleting retweets
3. Standardizing the dog ratings
4. Changing the timestamp to correct date time format.

5. Some of the numerators were actually decimals. Hence they had to be corrected.
6. The original dataset had three predictions for the dog types and one column for confidence interval. The dog type had to be reduced to one single column.