Twitter Data Wrangling Efforts

Firstly, I gathered data from 3 data sources, first one by downloading it from the Udacity site then began wrangling this twitter dataset by performing a visual assessment in which I browsed through the data to get acquainted with it and figure out which problems I'll need to solve, and see whether I was dealing with dirty or messy data. I then moved on to programmatically assessing the dataset.

The second dataset was downloaded using request and the third was data gathered from twitter API's and JSON

I began by importing the packages I would need to assess, analyse and visualize the data, namely, pandas, numpy, tweepy and requests etc.

I used pandas to read the data, then printed it by using .head to further look at the dataset in a table format for more assessments
I used more assessment tools such as {.info(), .isnull(), .tail(), sample(), .describe(), .duplicated(), .value_counts} to paint a clearer picture of the information in the dataset.

After viewing the data from the programmatic assessment, I documented the problems in bullet points under two headings, Data quality and Tidiness. The issues documented in the data quality section mainly highlight the following themes:

- Completeness – there was plenty of missing data
- Accuracy – two column names were unclear about what the data below was representing and the columns describing the types of dog had 'None' written in place of a null value or a zero and some names had question marks
- columns with the dog descriptions had to get merged – doggo, pupper, etc
- retweeted columns had to be removed
- column names to be changed and clearer
- Wanted to remove extra text and digits of timestamp and source columns
- Consistency – the rating denominator was inconsistent in some rows
- Changed datatypes of relevant columns
- Edited rating data with decimals and dropping columns that have numbers too large in the numerators
- There were redundant columns that were not necessary for this analysis and were removed
- The data under the following columns, 'timestamp' and 'source' had extra letters and numbers that needed to be tidied

The issues documented under Tidiness:

- The cleaned data frames needed to get merged
- Four columns have been combined into one that described the dog types

Before cleaning the data, I set a variable that will contain a copy of the edited clean dataset for further comparison during analysis to see the difference between the current dataset and the cleaned dataset.

I then moved on to resolving these issues by cleaning the data in which each step has been clearly defined, coded and tested for every bullet point problem I had documented when assessing the data.

I then visualized and analyzed the data and added final insights from the data