Data wrangling report

This project involved the data wrangling, analysis, visualization and reporting on the findings of analysing the WeRateDogs Twitter archive. The data wrangling process included gathering, assessing and cleaning the data as discussed below-:

1. Gathering data

The data was used in this project was gathered from three sources which are-:

a. The WeRateDogs Twitter Enhanced archive which was provided by Udacity and was manually downloaded to the local environment.
b. The Image Predictions data which was downloaded from the cloud hosted by Udacity using the Python module called request.
c. Lastly, we used the tweet_id in the first data frame to extract extra relevant tweets data. To extract data from the Twitter API a developer account was needed, which I signed up for. After getting the keys and tokens I initiliased them and together with the Python's Tweepy module I was able to fetch the necessary data.

2. Assessing data

After gathering all the data and storing it on the local environment, I read the data to dataframes using Python's pandas module. After that, I went on to assess the data using various Python's pandas methods like head(), info(), value_counts(), duplicated(), describe() and many more.

By using these methods, I was able to detect some quality and tidiness issues that my data frames had, such as incorrect datatypes, missing data, invalid data, duplicates, unnecessary columns etc.

3. Cleaning data

After I have assessed the data, I started cleaning the issues discovered in the Assessing data process. Some of the main data cleaning process I performed were-:

a. Changing data type of timestamps to datetime and IDs to strings.
b. Melting columns
c. Dropping columns which I fount not useful to my project
d. Changed all the 'None' values to Numpy's NaN values.
e. At last I combined the three data frames to one single data frame.