Report on Data Wrangling for WeRateDogs Twitter Archived Data

My data Wrangling process starts with Data Gathering. To gather the needed data, I downloaded the WeRateDog tweets and archived data directly from the Udacity website. Another data needed is the image prediction file I downloaded programmatically from a publicly accessible URL provided within the projects. The last step I took in my data gathering was to use Twitter API to access WeRateDogs tweets in other to gather additional columns needed for my analysis. Luckily, I already have access to the Twitter Developer account, so it was direct as no application was needed. I only reused my previous credentials. Tweepy Library was used to access the Twitter API programmatically to access tweets using the tweet_id and get extra tweets' metadata needed for the Wrangling Process. The request library was used to download the prediction file directly from provided URL. All gathered were then saved into file reusability.

Moving to the next step of the wrangling process, the datasets were loaded into different pandas DataFrame followed by an assessment of the gathered data using visual and programmatical approaches to identify twelve data quality issues and two tidiness issues. Some of these issues were discovered while trying to clean up other issues. To perform the visual assessment, I used a combination of Jupiter notebook and Microsft Excel to scroll through the dataset for visual assessment and used pandas and general python programming to access the dataset programmatically. The issues identified were documented for reference during the cleaning stage.

To clean and fix the identified issues, I used the Define, Code and Test approach for all Quality and Tidyness issues employing different functions from the panda's library and other general python functions. After cleaning the identified issues, the new dataset was saved in a separate file.

To wrap up the process, I identified three fundamental insights and presented a visualisation to show the tweets counts from different sources as indicated in the dataset.