Wrangle Report

This document is a brief wrangling data report from the wrangle_act.ipynb Python file. The aim of this paper is to show a summary of the wrangle efforts made in the WeRateDogs twitter dataset (https://twitter.com/dog_rates).

These wrangling efforts were divided in three steps described bellow:

1- Gathering Data:

In this step, the three dataframes were taken from WeRatedogs by three different methods:

1) A .csv file manually downloaded using pandas.read_csv command
Command documentation:
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

2) A .tsv file programmatically downloaded using the os and requests Python libraries
os librarydocumentation link: https://docs.python.org/3/library/os.html
requests library link: https://pypi.org/project/requests/

3) A json file downloaded from Python Twitter API - Tweepy (http://www.tweepy.org/)
Tweepy documentation: http://docs.tweepy.org/en/latest/

2- Assessing Data:

In this step the three dataframe files were assessed in order to discovery the issues according to the Four Data Quality Dimensions: **Completeness, Validity, Accuracy and Consistency**. An Udacity view from Data Assement Dimensions (since most papers consider 06 data assements dimentions https://smartbridge.com/data-done-right-6-dimensions-of-data-quality/)

3- Cleaning Data:

In this step the three data frames were merged into a one dataframe named df_twitter. I must highlight that only the data with tweet_ids in all the three dataframes were kept in the final df_twitter dataframe. Pandas commandas were used during this step in order to query, drop rows and columns and manually correct rating_numerator and rating_denominator values according to specified query conditions.