Wrangling Report

This report is to explain and summarize the process of data wrangling done in the "WeRateDogs" project requested by Udacity

Steps of successful data wrangling:

1. Process of gathering the data
2. Process of assessing the data
3. Process of cleaning the data

- ## Data Gathering Process:

  In the requested project, data was divided into 3 portions/sources. In order to conduct a successful gathering process, we must obtain all the data from the different sources.

  First source was a .csv file to be downloaded manually from the Udacity resources page. Next step was to correctly open this file in the Jupyter Notebook and read its contents in a Pandas DataFrame.

  Second source was a .tsv file to be downloaded programmatically from a given URL, using Python Requests library. Then read this file and construct a Pandas DataFrame.

  Third source was collecting certain data from Twitter account of "WeRateDogs" via Twitter API using Tweepy library and then storing it in a text file. Next step was to read that text file and store its data in a Pandas DataFrame. Unfortunately, this step required authentication from Twitter which was not granted, so I used the text file provided by Udacity.

- ## Data Assessment Process:

  The obtained data frames were then assessed in two steps. First step was visually inside a Jupyter Notebook with simply printing them using Pandas. Second step was a programmatic assessment made inside Jupyter with Pandas using the following functions, .info(), .head(), .sample() and .value_counts().

  The datasets were examined under two criteria, quality and tidiness. Quality refers to issues related to the content of the data, sometimes called dirty data. Quality dimensions are completeness, validity, accuracy, and consistency of the data under investigation. Tidiness refers to issues related to the structure of the data, sometimes called messy data. Tidiness dimensions are that each variable forms a column, each observation forms a row, and each type of observational unit forms a table. As soon as an issue was detected it was documented under whether quality or tidiness.

- ## Data Cleaning Process:

Cleaning process can be defined simply as solving or fixing the issues discovered in the collected data, programmatically, so that the data can be analyzed, giving the proper insights. Lack of cleanliness in data makes it so hard to perform analysis upon. The cleaning process is done in Jupyter Notebook and using Pandas, and it is conducted following the standard process of define, code and test for each of the issues, tackling them in a logical order. It is common to address quality issues first, then move to tidiness.

After completing the cleaning process and saving the final DataFrame into a .csv file, we are ready for Analysis phase.