Wrangle Report

Introduction:

The purpose of this project is to put into practice what I already learnt in the Data Wrangling Data Course that it is a part of Udacity Data Analysis Nanodegree program.

The dataset that I will wrangle is the tweet archive of twitter user WeRateDogs.

Project Details:

The tasks of this projects are:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

❖ Gathering Data:

We were required to gather the data from 3 different sources:

- Twitter Archive file:

  "The twitter-archive-enhanced.csv" is provided by Udacity, downloaded manually then was loaded from the CSV file into a Date frame

  This archive contains the basic tweet data (tweet ID, timstamp, text, etc) for all tweets as from August 1, 2017

- The tweet image predictions:

  This file contains the top predictions of dog breed for each image from the twitter archive. The dataset contains the top predictions, tweet ID, image URL, and the image number corresponding to the most confident prediction.

- Twitter API File:

  This contains the tweet ID, favorite count, retweet count. Data was provided by Udacity, downloaded manually then was loaded from the tweet-json.txt file into a pandas data frame

  Another way of obtaining the data would have been by requesting permission from twitter and using the python library called tweepy.

❖ Assessing Data:

Assessing Data describes inspecting the data visually and programmatically. **The part of visual inspection involves:** observing the structure of the data which is a major set-back when dealing with big data.

Quality has to do with the content of the data. This includes completeness, validity, accuracy, and consistency of a dataset.

❖ Cleaning Data:

The quality and tidiness issues identified in the Assessing Data section are cleaned:

This is regarded as the final step of the Data Wrangling Process. It contains the Define, Code and Test phase. Before providing solutions to this issues, I try to create a copy of these datasets to keep the original ones save, and programmatically ensured necessary python packages and libraries were used effectively.

I merged the copies into one Data Frame. I found the solutions to the various quality and tidiness issues and tested by checking to see the solutions were properly implemented.

Finally, I assigned the cleaned Data Frame as 'twitter_archive_master'.