Report on the Data Wrangle project

As part of an assessment toward the completion of Data analyst nano Alx project, I was required to complete a Data Wrangling and Analysis project.
This report showcase the steps and procedures I undertook to clean up the dataset, analyzed and visualized it.

The dataset that I worked on is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This is a Twitter account that rates people's dogs with a humorous comment about the dog.

The project was completed on my local computer(outside udacity project work space) and I ensured that I adhere to the software requirements as stipulated in the project environment requirements.

The Wrangling process is divided into three stages:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

The data was gathered from three different sources

- Enhanced Twitter Archive: This data was downloaded manually from the given link and was uploaded to the local project wok space on my computer. It contains information like Dog names,rating, dog stage etc.
- Image Prediction File: I made used of Requests Library to download this file from the Udacity Servers. This file consists of tweet_id, image url and 3 image predictions with their confident predictions.
- Twitter Api: I was able to query Twitters Api to obtain favorite_counts,retweets_counts with tweet_id obtained from Enhanced Twitter Archive, I had to apply for Twitter developer account on the twitter platform to be able to gether from twitter Api.

Assessing Data

I assessed the 3 data on two levels: visually and programmatically to be able to pinpoint some data qualities and tidiness issues that were associated with the data.

A) **Tidiness:**

- doggo, floofer, pupper, puppo columns in twitter_archive_enhanced dataset should be combined into a single column
- Related Datasets that exists separately

B) **Quality**

From tweet image predictions Dataset

- underscore used to separate some predicted dog names across p1, p2 and p3 columns

- Some dog names started with lower case while some start with upper case
- Wrong data type ascribe to tweet_id column int64 instead of string

From Twitter_Api Dataset
- Wrong data type assign to the id column int64 instead of string

From Twitter Archive Data

- Wrong data type assign to the timestamp column (object instead of Datetime)

- Wrong data type assign to the tweet_id column (int64 instead of string)

- There are 181 retweet values as shown by the retweet_id
- some rows have rating_denominator values not equal to 10¶
- Invalid values in name column

Cleaning Data

The following operations were done to get rid of quality and tidiness issues raised when assessing the Data

- Renaming of columns
- Extraction of values from text
- Changing Data types to appropriate ones
- Merging of Datasets
- Dropping Columns with high missing values