Wrangle Report

Introduction:

The purpose of this project is to practice what I learned in data wrangling from Udacity Data analysis professional track, the dataset that is wrangled is tweet archive of twitter user @dog_rates, known as WeRateDogs. WeRateDogs is a twitter account that rates people's dogs with funny comments about the dog, the rating always have a denominator of 10.

This report describes my wrangling efforts.

Project wrangling steps:

- **Gathering data**
- **Assessing data**
- **Cleaning data**

- Gathering Data

  The data consist of three datasets:

  - Twitter Archive file: provided by Udacity as .csv and download it manually.
  - Tweet image prediction: provided by Udacity server as .tsv and download it programmatically by using provided url and request library.
  - Twitter API: I used tweet_json.txt file read file line by line into list and create Data Frame from the previous list that contains the id,retweet_count, fav_count

- Assessing Data: by 2 methods

  Visually: by using print the data frame to check the data quickly and by open it by excel

  Programmatically: by using different methods like (.head(), .info(), .sample(), etc..).

And separated the issues to Quality issue and Tidiness issues.

- Cleaning Data:
  Divide the cleaning into 3 parts: Define, Code, Test the code, on each step the issue is described and assessed.

  First, Copy of the original data frame, due to if there was an error, I could create new copy from the original data frame.

  The challenging cleaning step was names of dogs are invalid ('None, a, an instead of name') I used replacing them with NaN.

- Conclusion

  Data wrangle is fundamental skill whatever handling the data you should be familiar with

  I used Python programming language and its packages, there are several advantages of Python (much better than Excel)

  It can deal with unstructured data such as Json (tweets) or Structured.