Wrangle Report

Introduction: The purpose of this project is to wrangling data section from Udacity Dataset. The dataset that is wrangled (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates dogs .These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "**they're good dogs Brent**."

Project details:

The tasks of this project are as follows:

- Gathering data:
  Gathering data is the first step in data wrangling. Before gathering, we have no data, and after it, we do. Depending on the source of your data, and what format it's in, the steps in gathering data vary.High-level gathering process: obtaining data (downloading a file from the internet, scraping a web page, querying an API, etc.) and importing that data into your programming environment (e.g., Jupyter Notebook).
  I gathered data from 3 different recourses
  1-**csv file(twitter_enhanced_archive.csv)** Was given from udacity
  2-**tsv file(image-predictions.tsv)** This file is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and url and it contains what breed of dog (or other object, animal, etc.)
  3-**API & json.txt** read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count
  file In my case I did not download the data from the API I get the tweet_jason.text from udacity

- Assessing data:

  assessing data is the second step in data wrangling. When assessing, you're like a detective at work, inspecting your dataset for two things: data quality issues (*i.e. content issues*) and lack of tidiness (*i.e. structural issues*).

  1-Quality

  qualit : issues with content. Low quality data is also known as dirty data.
  *Completeness, validity, accuracy, consistency (content issues)

  twitter-archive-enhanced

  - some dogs name are invalid
  - Delete columns that won't be used for analysis
  - Correct denominators other than 10
  - Decimal Dog Rating
  - timestamp is 'str' instaed of 'datetime'
  - tweet_id is 'str' instaed of 'int'

  image-predictions

  - delete 66 jpg_url duplicated
  - some p names start with lowercase and other not
  - some p names contains _ instead of space
  - Delete columns that won't be used for analysis
  - tweet_id is 'str' instaed of 'int'

  tweet-json

  - tweet_id is 'str' instaed of 'int'

  2-Tidiness

  tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements: Each variable forms a column. Each observation forms a row. Each type of observational unit forms a table.

    - All tables should be part of one dataset
    - four columns (doggo, floofer, pupper, and puppo) should be one
- Cleaning data:

  Cleaning your data is the third step in data wrangling. It is where you fix the quality and tidiness issues that you identified in the assess step. In this lesson, you'll clean all of the issues you identified in Lesson 3 using Python and pandas.
- Visualization and Analyze Data

  using Visualization show how does the cleaning effect on the data and make it easier to compare or show the relations between the elements in the dataset