wrangle_report

Notebook:	Udacity		
Created:	28-Sep-19 9:34 AM	**Updated:**	28-Sep-19 10:15 AM
Author:	prashant kumar		
URL:	http://localhost:8888/notebooks/Desktop/City/project4/mine/wrangle_act.ipynb#		

Wrangle and Analyze Data
Udacity Nano Degree Program

Introduction
The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs rate's pictures of people's dogs in a humorous manner, most often giving ratings higher than 10/10.This project briefly describes about my wrangling efforts

Project Details

- The tasks performed in this project include:-
    - **Gathering Data**
        - **WeRateDogs Twitter archive:-** This file can be downloaded from the link given in the portal
        - **Tweet image predictions:-** Indicates what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
        - **Tweet's JSON Data:-** Using the tweet IDs in the WeRateDogs Twitter archive, we can query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.
    - **Assessing Data**
        - Once the the data is gathered we can assess the data in two ways:-
            - **Visual Assessment:-**In this process we put an eye on the gathered data and get overall overview.For this purpose I opened the Csv file in the excel sheet and also made use of jupyter notebook to see different parts of the data
            - **Programmatic Assessment:-**In this part I made use of different pandas methods to analyse the data such as info,describe,sum,value_counts,shape,head etc
        - After assessing the issue I commented about the Quality and tidiness of the data which can be seen in Jupyter Notebook.
    - **Cleaning Data**
        - This is the final stage of data wrangling.This stage consists of 3 parts
            - **Define:-**Here I defined the cleaning step which is to be performed.For example If we need to drop some unnecessary columns ,we can mention here.In this project different definitions include:-

        - Merging all the 3 dataframes
        - Combining dog stages into one column.
        - Replacing names that doesn't look realistic.
        - Dropping the unwanted columns.
    -
    - **Code:-**Here,we write the actual code with performs the action written in the defined part.For example here we write code to drop the columns.
    - **Test:-**Finally in this part we check whether the above written code works as per our requirement i.e we check whether those columns have been dropped or not for which we have written the code.
- **Conclusion:-**
    - Data wrangling is the backbone of the data analysis process.For this purpose there are several python libraries which has been used such as numpy,pandas,matplotlib etc.After completing the wrangling process several meaning observations can be seen.