Title: Report on Data Wrangling steps for We Rate Dogs Twitter Dataset

Abstract: The following report shows a detailed description of the data wrangling process starting from data gathering, Data assessing, Data cleaning process.

Data Gathering:

This is the first step of any data wrangling process which involves to gather the raw data from different data sources. In my experiment I had the opportunity to gather data from different source format and transforming them to create pandas dataframe.

- **Step 1:** Here I have been provided with a .csv file namely **twitter-archive-enhanced.csv** .This file has been transformed to a pandas dataframe.
- **Step 2:** Next comes with a description of an URL:'*https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv*'. The file with an extension of. tsv has been downloaded via Request library with the Get method. After downloading the following file has been opened with read-write format and stored in a pandas dataframe.
- **Step 3:** At the next step I have gathered additional data from Tweepy API with proper authentication (consumer_key,consumer_token,access_token,access_secret).
- **Step 4:** Upon successfully extracting data from tweepy, I have stored the end file into a .txt file namely 'tweet_json.txt'.This is to be noted that all the data has been stored as a json text.
- **Step 5:** The data gathered from Step 4, has now been loaded with json library function in python.Each of the json text line has been analyzed and being inserted in an empty dataframe.

Data Assessing:

After successfully gathering all the required data into pandas dataframe, it's time to assess them for the next cleaning process.The source dataset thus I have gathered is :

1. twitter_archive (dataframe to host all the archival data)
2. image_predictions(dataframe which contains predictions of dog images)
3. tweets_data(dataframe containing favorite counts and retweet counts)

- **Step 1:** Each of all the dataframes involved to view a sample of the dataset visually and programmatically.
- **Step 2:** Check for the duplicacy of entries in the entire dataframes.
- **Step 3:** Check for the data type inconsistency with respect to the columns.
- **Step 4:** Check for any missing values within the dataframes.
- **Step 5:** The probabilities associated with the dataframe 'image_predictions' are being tested if any of these greater than '1' or less than '0'.

- **Step 6:**Any of the quality issues and tidiness issues found in above dataframes are being noted for the data cleaning process.

Data Cleaning:

Once I have gathered and assess all the dataframes, data needs to pass through cleaning process in order to reduce quality and tidiness issues.This is one of the important steps to carry out before making visualization,predictions etc.

- **Step 1:** The first step is to take copies of all the dataframes before starting with any cleaning process.Data must be cleaned on the copied dataframes instead of the original ones.
- **Step 2:** Perform the required modification which has been highlighted in assessing phase.
- **Step 3:** Cleaning must be processed through each dataframes.
- **Step 4:** Once completed with probable cleaning solutions, data must be passed for subsequent phases.

Conclusion:

Data Wrangling effort varies from person to person as it totally depends how one sees data,the quality issues ,tidiness issues are the one which may change from one's perspective.The associated project doesn't guarantee omission of all kinds of quality and tidiness issues.