Wrangle Report

This Project was based on Data Wrangling i.e. Data Gathering, Data Assessing and finally Data Cleaning. The programming language and IDE used was **Python** and **Jupyter Notebook**.

The main objective of this project was to wrangle and analyse the "**WeRateDogs**" twitter account tweets. WeRateDogs rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. WeRateDogs downloaded their Twitter archive and sent it to Udacity. This archive contained basic tweet data (tweetID, timestamp, text, etc.)

The first task was **Data Gathering** and mainly divided into three parts:-

1. The first data set i.e. **Twitter Archive** has already been sent to Udacity.
2. The tweetID's obtained from the above data set was used to query the Twitter API and collect additional data for the WeRateDogs tweets. **Tweepy** library was used to connect with Twitter Developer account and query the Twitter API.
3. Using python's request library, the corresponding **Image Predictions** data set was downloaded.

The next task was **Data Assessing** and was achieved through Python programming. The main purpose of this task was to identify data quality and tidiness issues.

1. **Twitter Archive**- Five data quality issues and four tidiness issues were found in this data set. For example retweets in the Data Frame needs to be removed (as only original tweets with images are required for analysis), erroneous data types for tweetID, erroneous dog ratings, date and time column segregation etc.
2. **Image Predictions**- Three data quality issues were found. For example erroneous datatypes, column renaming and data consistency (mainly maintaining same format for all the data present in the column).
3. **Tweets Json**- This data set was more or less cleaned.

The third task dealt with **Data Cleaning** for the issues that were identified during the Data Assessment. **Pandas** library was mainly used for Data Cleaning. Methods like **drop()**, **info()**, **describe()**, **astype()**, **iloc()** etc were used for this purpose. **For** loop, **if else** condition was also used to complete task.

Data Storing was followed by Data Cleaning. The three cleaned data sets were stored in **csv** format. Important columns from these individual data sets were picked and then combined to form one single master data set.

Finally the project was ended with **Data Visualization** obtained from the above created master data set. The cleaned data set was analysed rigorously and different insights were drawn from it like **Retweet_Count**, **Favorite_Count** per tweet, most appearing **dog names** etc. All these inferences were plotted using python **seaborn** library.