Wrangling Act

1. Overview:

The project focuses on gathering, assessing , cleaning and analysing data from the tweets from the user id @dog_rates . WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. In this project we focus on cleaning the data gathered from three sources and analysing it to draw out conclusions.

2. Components of Wrangling:

The Wrangling process focuses on three main segments:

I. Gathering

II. Assessing

III. Cleaning

All these parts of the Wrangling act for the given data will be explained sequentially.

I. Gathering

In the first step of the Wrangling Act I gathered data from three different source , from where the data is cleaned and merged eventually.

Sources are:

a) The WeRateDogs Twitter archive is provided by twitter_archive_enhanced.csv which is downloaded manually .

b) The next table to be extracted is the Image Prediction table which I downloaded programmatically. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library.

c) The next piece of data was the most difficult to gather out of the three. Using the tweet IDs in the WeRateDogs Twitter archive, 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. The data in this table consists of information like retweets and favourites for that particular tweet. It took me several tries and hours for wait to get finally data

II. *Assessing:* After gathering all the data I divided the assessing part into two segments:

a) Manual Assessment b) Programmatic Assessment.

a) For Manual Assessment, I looked for errors by looking from the tables visually and check for Data Tidiness.

b) For Programmatic Assessment , I looked the .info() function to check for Data Quality . I also assessed individual columns in seek for any data issues.

At the end of this process I came across 9 Data Quality Issues and 3 tidiness Issues

III. Cleaning:

For the final step of the wrangling act I rectified 9 Data Quality Issues and 3 Tidiness IssuesI initially made copies of the original tables to allow me to work with the copied table and experiment with them which turned out to be extremely useful as I lost the data in the copied data several time due to wrong code or logic.

Next I rectified the above mentioned errors one by one, starting of with the tweet_additonal tables which contained data on likes and retweets and rectified the three issues I noticed in the table to enable better final analysis.I followed the same procedure for the next two tables, images_df and and df which contained information about image prediction and Tweet archives respectively.

After fixing all the quality issues, I moved forward to solve the tidiness issue. I extracted the data from the four tables of dog stages(doggo,floofer,puppo,fluffer) to single column of Stages. Additionally I also changed the data type of `timestamp` column from object to datetime and extracted the date value from it to make a separate column for date.

Finally I merged all the three tables into a single master table to pave way for further analysis.

IV. Storing:

For the final step of Wrangling, I stored the table into a .csv file .

Inference: Data Wrangling is an important procedure to be followed by dealing with data. It can be tough, tedious and tiresome but it makes the following steps a lot easier. Cleaning a messy and untidy data is extremely important and Data Wrangling is the tool which paves way for solving these issues.