Wrangle Report

In this project, I am applying my data wrangling skill learned from Udacity Data Analyst Nanodegree. The dataset that I will be wrangling is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. 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.

This Wrangle Report briefly describes my wrangling efforts with the purpose of having a dataframe that is clean and ready to be used for analysis and visualization.

Wrangling exercise consists of three main steps:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

There are 3 pieces of data that I gathered from a variety of sources and in a variety of formats:

1. The WeRateDogs Twitter archive given by the instructor in CSV format. I used Pandas read_csv function to read the CSV file into the notebook.
2. The tweet image predictions, i.e., 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 was downloaded programmatically using the Requests library and the URL given.
3. Each tweet's retweet count and favorite ("like") count from the Twitter API that I queried using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. I read the text file line by line to extract the tweet id, favorite count, and retweet count and store it into a Pandas dataframe.

Assessing data

After gathering all 3 pieces of data and reading them into the notebook, I did both visual and programmatical assessment to identify any Quality (content-related) and Tidiness (structure-related) issues in the 3 dataframes. For the visual assessment, I displayed 10 sample rows from each dataframe in the notebook and also open the files in Excel and investigate on each row and column to detect issues. For the programmatical assessment, I used Pandas function, such as info, duplicated, value counts.

I noted down 9 Quality issues and 3 Tidiness issues that I will clean in the next section.

Cleaning data

The first step in doing the cleaning step is to create copies of each of the original dataframe, so if there is any error, the original data is not lost or broken.

The cleaning section is divided into three sub-sections: define, code, and test. Each issue cleaning has its own define, code, and test sub-sections in the notebook.

This step is the most challenging and time-consuming part for me because not all of the issues could be cleaned easily. I had to search for the methods from other sources, such as Stack Overflow.

One of the hardest issue to clean was the rating denominator column, because not all rows have 10 as the denominator, which is supposed to be the base. I had to do a 2-part cleaning for this issue to get the result that I wanted.

Another hard issue to fix was the rating numerator column because I was still not skillful in writing RegEx and also I needed to write a complex loop of code. It took me some time to get this corrected, but this cleaning process taught me a lot.

Conclusion

I really enjoyed doing this project because I got to learn a lot about gathering data from different sources and in different formats. The cleaning process also got me thinking hard and it was a challenging, yet fun exercise. Looking at the clean data at the end of the cleaning process was very worth the time and brainpower I poured out for this project.