This data wrangling and analyzing project consists of three datasets. The WeRateDogs Twitter archive is given in csv format, the tweet image prediction is in tsv format. We can easily import these two from python using pd.read\_csv(). The third dataset is tricky. We need to query the Twitter API to get each tweet’s retweet and favorite count. This is really new and challenge to me. Before this course, I have never done such similar projects. Query Twitter API needs a twitter developer account. It seems a high requirement for your coding skills. Fortunately, the project is well explained every step. It was not hard to register a developer account. Then the next step is querying the API and storing the tweet data in JSON format. At the beginning, I was very confused why we need to form the data in JSON format, convert to a text file, and then read the text file line by line to python data frame. It seems a redundant procedure. After a lot of searching on the project Q&A section, I understand that this is not a necessary procedure, but rather a good practice. I was very new to the JSON format, this is a lot of learning through website and Youtube videos, but I know the learning process is also a part of the goal of the project. Even an experienced programmer can’t know everything before work. Faster learning by self is an important skill for candidate. After about five hours effort, finally, I gathered all three datasets into three data frames in python.

The second step is assessing. Firstly, I did visual assessment for all three datasets in Excel. I noticed that the expanded\_url column for the twitter\_archive table sometimes has two urls in one record. That is a tidiness issue. Also, I noticed there are lots of missing value in in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_user\_id and retweeted\_status\_timestamp columns in twitter\_archive table. That’s a quality issue.

Secondly, I use programmatic assessment. I found the data type for all ‘tweet\_id’ column in three datasets were wrong, it should be a string not number. This is a quality issue. Also, there are some other data type misused for other column when I use table.info() command. In order to make sure an effective analysis, I also think the rating denominator and numerator should combine to one column as rating. And the different stage for the dog should combine to one column for easy analysis.

After I listed all the issues, next step is cleaning. It didn’t take me much time to find a way to clean the data, but since my python skill is not excellent. I always need to search the exact command before I wrote code, that’s time consuming, which makes me think I need to take a python course to polish my python skill.

Next step after cleaning the data is analysis. This dataset is all about dog, unfortunately, I never have a dog in my life. I am not very familiar for the terms of dogs, such as what stage mean, what breed means. It took me some time to google and understand these terms before I can do any analysis.

This project provides me a good opportunity to use different ways to wrangle data. Especially scrapping data from website, which I have never practiced before. Now I am very interested in this skills, and I am going to query some other website to complete the project I am interested in.