Wrangle Report

by Nicholas Giddings

The following report documents my steps taken during the twitter data wrangling project.

After downloading the required documents from Udacity as well as using the twitter API to download the image_predictions.tsv I opened them up on my terminal. I wanted to work with a single complete dataframe for my wrangling working and immediately noticed a couple of issues that needed to be cleaned before I could join the three tables.

1. json_df does not contain a column that can be used for joining tables.
2. json_df contains many unnecessary columns.
3. Each dataframe has a different total number of rows therefore some rows will have missing

So firstly I renamed the id column to tweet_id so that I would have a common column to join my dataframes then I cleaned up the json_df and removed the extra columns that weren't necessary for the analysis. After that, I combined all three tables using a left inner join onto the archive file so that any rows that did not contain the data we needed would be automatically removed.

With those three issue sorted, I went ahead and created a new .csv file of the complete dataset. I then used a variety of functions to assess the state of the data. Some of the functions I used were:

- .sample
- .head
- .tail
- .info
- .describe
- .value_counts
- .query

I noticed quite a few issues but for the purposes of this project I tried to limit it to a few that I thought would be important for analysis. I noticed the following quality issues:

1. json_df does not contain a column that can be used for joining tables. (Completed)
2. json_df contains many unnecessary columns.(Completed)
3. Each dataframe has a different total number of rows therefore some rows will have missing values. (Completed)
4. Contains 70 retweets.
5. Contains 23 replies.
6. Timestamp is not datetime format.
7. All tweets after August 1st, 2017 should be removed.
8. Numerator and Denominator sometimes takes the wrong information.
9. Ratings numerators and denominators have outliers.
10. Extra columns that are not needed for analysis should be removed.

I also noted the following tidiness issues:

11. All three dataframes should be joined into a single dataframe as each row represents a single tweet. (Completed)
12. Dog stages should be converted into a single columns with original columns as values.

Issue: Contains 70 retweets & Contains 23 replies.

I noted that any rows that had data in retweeted_status_id and in_reply_to_status_id were either replies and retweets which were easy to clean by only keeping values that contained nulls in these fields.

Issue: Timestamp is not datetime format.

Another easy to clean issue with pandas to_datetime function

Issue: All tweets after August 1st, 2017 should be removed.

On checking, I only noticed 2 tweets which I removed easily.

Issue: Numerator and Denominator Errors.

This was the most time consuming issue as I had to manually go through lines of information. On final reflection, It probably would have been ok to drop these lines for the sake of efficiency. To clean this, I went through each row that had an issue and manually replaced the data. With every iteration, I noticed further issues with ratings so in the took a lot of time.

Issue: Dog types in multiple columns.

This one was also challenging, at first I tried to use lambda but had a lot of difficulties with that so I went for a simpler way of just replacing ‘None’ with blanks then combined all the data into a single column, changed rows with double entries and dropped the old columns.

Finally I erased all the extra columns that wouldn’t be needed for analysis and saved the final document into a cleaned complete dataset.