We Rate Dogs: Wrangling Process

In order to draw conclusions, insights, and create visuals, diverse sets of data needed to be collected from several locations, cleaned, and organized before they could be merged together into one data frame to be analyzed. In the following text I will describe that process.

The first step was to collect the data from three different locations. First a “local” csv file named twitter-archive-enhanced.csv needed to be imported from the virtual environment contained in the jupyter notebook cloud. That was a simple process of using pandas built in import function read\_csv. The second data set was contained ina tsv file that was hosted on a website. For that I was able to utilize python’s built in request function that uses a standard http GET request to import the file. The final dataset was contained in twitters database which can be accessed through their custom api. That involved creating an account, using the keys associated with the account to initialize an api object. In this particular case, I needed to use their version 2 and initialize a client object, as the api object does not allow for querying the specified data without elevated access, but basic free access was available through the client object. I created a list of tweet id’s from the other two datasets to query the correct tweets from the twitter api and I was able to extract tweet\_id', retweet\_count, and favorite\_count.

Then next step was to assess and clean. I used various pandas functions to inspect the data tables to see what was included and look for pitfalls. This involved inspecting the entire tables as a whole as well as using functions like .info(), value\_counts() and .describe() to get a look at datatypes, averages, minimums and maximums as well as the number of occurrences of different values in different columns. Here is a list of the quality issues I found and corrected:

1. Cast timestamp column to datetime datatype.
2. Change denominators to match the text.
3. Cast pupper, doggo, and floofer to category data types.
4. Change all dog breeds to lower case.
5. Turn rows with string "None" to NaN.
6. Drop retweets from table.
7. Cast numerators and denominators to float
8. Cast id columns to strings from ints.

Since data wrangling is an iterative process, I actually fixed one more quality issue while I was trying to create my insights.

9. Remove incorrect non-dog names.

The two tidiness action points were the following:

1. Melt dog stage columns (doggo, fluffer, pupper, puppo).
2. Merge the three datasets

As with cleaning, I ended up adding one more organizational change to help in the insight process:

3. Add a column with the ratings calculated as a float by dividing the numerator column by the denominator column.

Once the data was clean, tidy, and well organized into one dataframe, I was able to start exploring statistical trends. I utilized built in functions such as .query(), .mean(), .unique(), .describe() to make calculations and inspect specific rows in specific columns. Finally, I was able to utilize matplotlib, to create visuals from my insights.