WeRateDogs Twitter Archive: Data Wrangling

To fully wrangle the data necessary for this project, I completed three main steps: (1) gathering, (2) assessing, and (1) cleaning data.

1. Gathering Data

The required dataset came from three sources: an on-hand file of the WeRateDogs Twitter archive, a URL for the image predictions data, and the Twitter API for additional tweet data. I gathered the first piece of data from the on-hand file by simply reading it into a DataFrame object. For the second piece of data, I retrieved the data from the URL using the Requests library, then wrote it into a tsv file, from which I was able to create a second DataFrame. For the third and final piece of data, I had to create a Twitter developer account to get access to the Twitter API. This was done by creating an API object through the Tweepy library which was authenticated using access tokens provided by my Twitter developer application. Once the API object was created, I used the tweet IDs from the archive file to query the JSON data of each tweet and wrote it all into a .txt file line by line using the json library. Then I parsed this file for the ids, retweet counts, and favorite counts of each tweet so that I could create another DataFrame. At the end of this step, I had three DataFrames containing data from each of the three sources.

2. Assessing Data

In this step, I visually and programmatically assessed my gathered data for issues in quality and tidiness. The main functions I used for this step were: .info(), .unique(), .nunique(), .value_counts(), shape, describe, and head() among others. I documented a summary of all issues in the notebook.

3. Cleaning Data

The first thing I did, once all the data was assessed for issues, was merge all three of the DataFrames into one master DataFrame since all of it pertained to one observational unit. I also created a copy of this master table to clean so that I had the original data to refer back to if needed. Then, I went down the list of quality and tidiness issues I created in the Assess step and addressed each one by one, starting with the tidiness issues and followed by the quality issues. For each issue, I defined the cleaning process, coded it, then tested it. Throughout this process I modified my assessment list with new observations. After all the cleaning was complete, I stored the final table into a csv file.