WRANGLE REPORT

Adrien Viani

This report will describe my data wrangling efforts on the WeRateDogs Project for Udacity. Data wrangling consists of 3 processes: gathering data, assessing data, and cleaning data

Gathering the Data

The data was gathered from three different sources and required parsing different filetypes programmatically to load the data into data frames. These files are described below:

1) *Twitter_archive_enhanced.csv* : This file was manually downloaded by clicking a link provided by Udacity. The data was then ingested using the standard read_csv() function

2) *Image_predictions.tsv* : This data was extracted programmatically using the python Requests library to send an html request to a URL provided by Udacity, shown below. The file was then read in using the read_csv method.

   https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3) *Tweet_json.txt* : The final file was built using Twitter's API, tweepy, to store each tweet as a line of JSON data that included at a minimum: Tweet Id, retweet count, and favorite count. This file was ingested line by line into a python list, and then the dictionary like structure exploited to create the necessary dataframe.

Assessing the Data

Various methods were used to assess the dataframes

Pandas.info(), to understand variable encodings, names, and dataframe size

Pandas.sample(), to quickly and easily display various rows of the data and visually inspect each entry

Regex, to parse through and identify the scope of text issues that were identified visually

isnull() to explore nulls and determine if they were null for a reason

.value_counts() to build aggregate counts of values and identify quality issues

From these methods, I was able to identify the following quality and tidiness issues.

Quality

1. The archive and json datasets should have no posts with retweets

2. Their should be no tweets with no images

3. Their are clearly incorrect dog names (a, the, an, officially, etc.)

4. id should be a string across all three data frames to avoid improper arithmetic operations

5. timestamp is not a date time object

6. tweet sources have complicated href strings that are difficult to make sense of

7. Decimal ratings in numerator / denominator are incorrectly formatted

8. Extra characters after ampersands in certain text row of the archive df

Tidiness

1. The dog stages should be consolidated to one column

2. Each of the different data files should be merged together

After identification, I defined a specific programmatic plan to clean the data. These are outlined below.

Quality (Definition of Solutions)

1. Delete Retweets

2. Remove posts without expanded urls

3. Determine names with only lowercase values, then check to see if "named" or "name is" in the text field of the post to determine the actual name. Replace with the actual name if available, otherwise replace with non

4. change each id variable type to string using astype(str) to recast

5. change the format to datetime

6. utilize regez to identify the source at the end of the href code that lies between the string >(.+_<

7. right a regex to identify decimal ratings in the tweet text, then either manually change the values or create a loop if the # is too large

8. for affected rows replace '&amp;' with '&' using string replacement

Tidiness (Definition of Solutions)

1. use melt to merge the columns and transform the "wide" columns to a long format, remove duplicates afterwards

2. use pd.merge() to merge the files using the appropriate left joins

Store Data

After addressing the issues, I stored the merged dataset in the 'Twitter_Archive_Master.csv'