Data Wrangling Report

Project objectives

The objectives of the project were:

1. Perform data wrangling (i.e. gathering, assessing and cleaning) on the three (3) provided data sources.
2. Store, analyze, and visualize the wrangled data.
3. Provide a written report on (1) Data wrangling efforts and (2) Data analyses and visualizations.

Step 1: Data Gathering

This step involved gathering the datasets together and loading them into pandas dataframes:

1. The WeRateDogs Twitter archive. This was manually downloaded from the link provided on udacity
2. The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the python's requests library from the provided URL.
3. Each tweet's entire set of JSON data in a file called 'tweet_json.txt' were stored using Twitter API and Python's Tweepy library. Then, each tweet's JSON data was written to its own line.

Step 2 : Assessing and Cleaning Data

After gathering all three pieces of data, they are then assessed visually and programmatically for quality and tidiness issues. The table below shows the observations in terms of quality and tidiness issues and the actions taken to clean the data

Quality

Dataset	Observation	Solution
twitter_archive	**timestamp** is string and should be datetime. **tweet_id** has type int64 and should be string	Change the variable type to datetime. **tweet_id** was changed to type string for uniformity
	In **rating_denominator** there are 15 unique denumerators, the key trend is that they are all in multiples of 10, except only 3	The 3 denominators (11, 7, 2) that are not in multiples of 10 were approximated to multiples of 10. Ie. (11=>10, 7=>10 and 2=>10)
	Missing values in several columns. Most of these columns are irrelevant to our analysis	These columns were dropped from the dataframe
	doggo, floofer, pupper, puppo columns contain 'None' value where NaN should be used. There are a few cases, where a dog has more than one style.	All 'None' values were changed to NaN. Multiplied dog styles were resolved during dataset tidying process and the logic described in the accompanying Jupyter notebook.

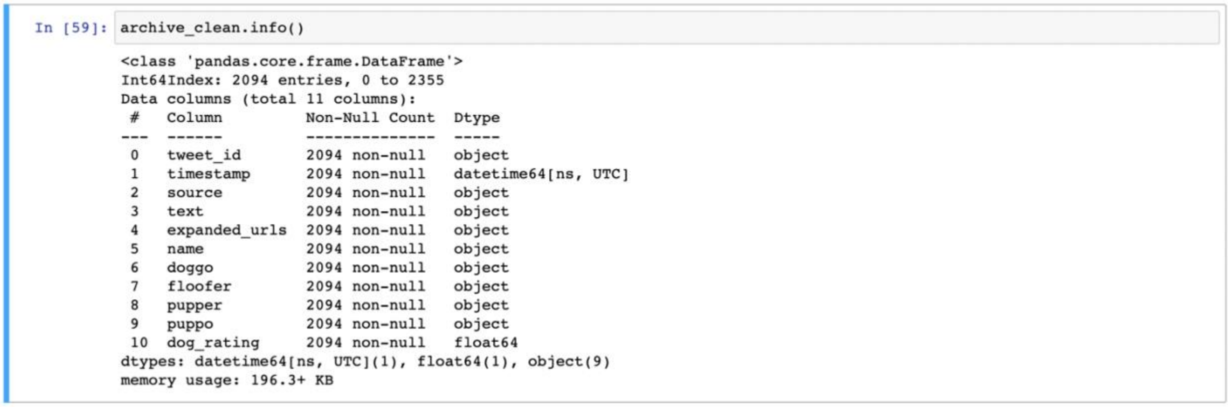
	There could be encoding problem for tweet_id = 668528771708952576 (the **name** value uses non-English characters).	The problem was noticed during review in Excel. In pandas dataframe, the encoding seems to be correct. No action taken.
	jpg_url contains two different path patterns to jpg files. This seems not to have any impact.	No action taken.
image_predictions	**tweet_id** has type int64 and should be string	**tweet_id** was changed to type string for uniformity
	The types of dogs in columns **p1**, **p2,** and **p3** had a mix of uppercase and lowercase letters. Need to change to lowercase	All names in columns **p1**, **p2**, and **p3** were changed to lowercase
tweet_json	The column **'id_str'** should be changed to **'tweet_id'** so it can be consistent with tweeter_archive and image_prediction dataframes	Column name **‘id_str’** was renamed to **‘tweet_id’**
all	Ensure only tweets before August 1st, 2016 were used	dataset.No tweet found beyond August 1st, 2016. Thus no action taken here

Tidiness

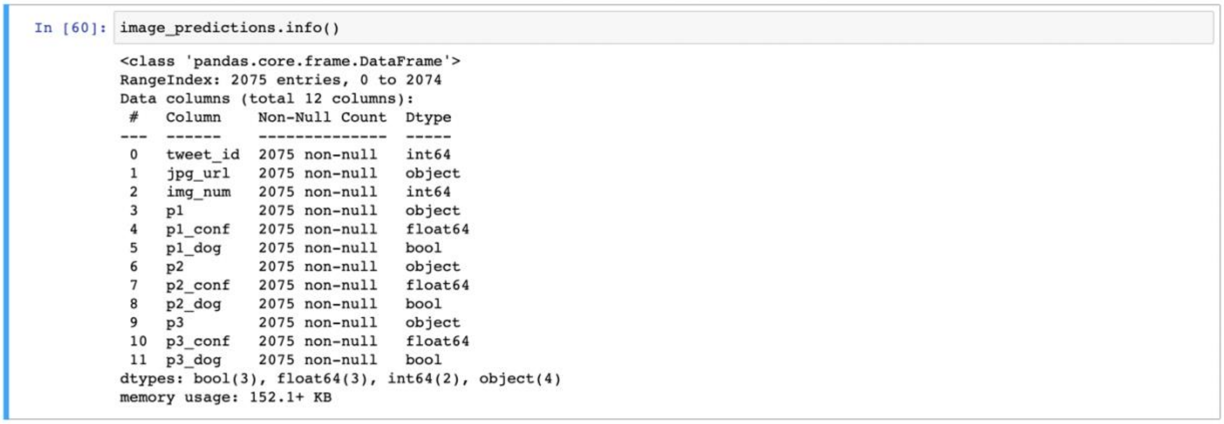
Dataset	Observation	Solution
twitter_archive	There are retweets and replies included in the dataset	Removed as per one of the project’s requirements.
	The source column in twitter_archive table looks messy and clutters the table.	Regex is used to extract source from the text avalaile
all	There are too many datasets and their overall structure is untidy.	One master dataset is created from the merger of the 3 datasets

Result

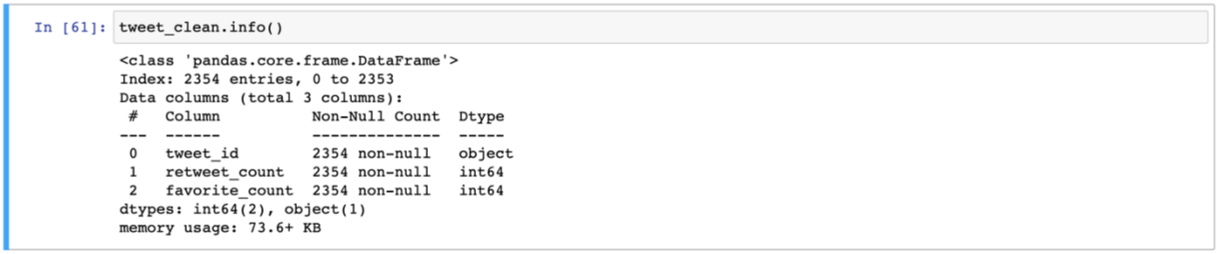
The figure below shows the final output after wrangling:



```
In [59]: archive_clean.info()

         <class 'pandas.core.frame.DataFrame'>
         Int64Index: 2094 entries, 0 to 2355
         Data columns (total 11 columns):
          #   Column         Non-Null Count  Dtype
         ---  ------         --------------  -----
          0   tweet_id       2094 non-null   object
          1   timestamp      2094 non-null   datetime64[ns, UTC]
          2   source         2094 non-null   object
          3   text           2094 non-null   object
          4   expanded_urls  2094 non-null   object
          5   name           2094 non-null   object
          6   doggo          2094 non-null   object
          7   floofer        2094 non-null   object
          8   pupper         2094 non-null   object
          9   puppo          2094 non-null   object
          10  dog_rating     2094 non-null   float64
         dtypes: datetime64[ns, UTC](1), float64(1), object(9)
         memory usage: 196.3+ KB
```

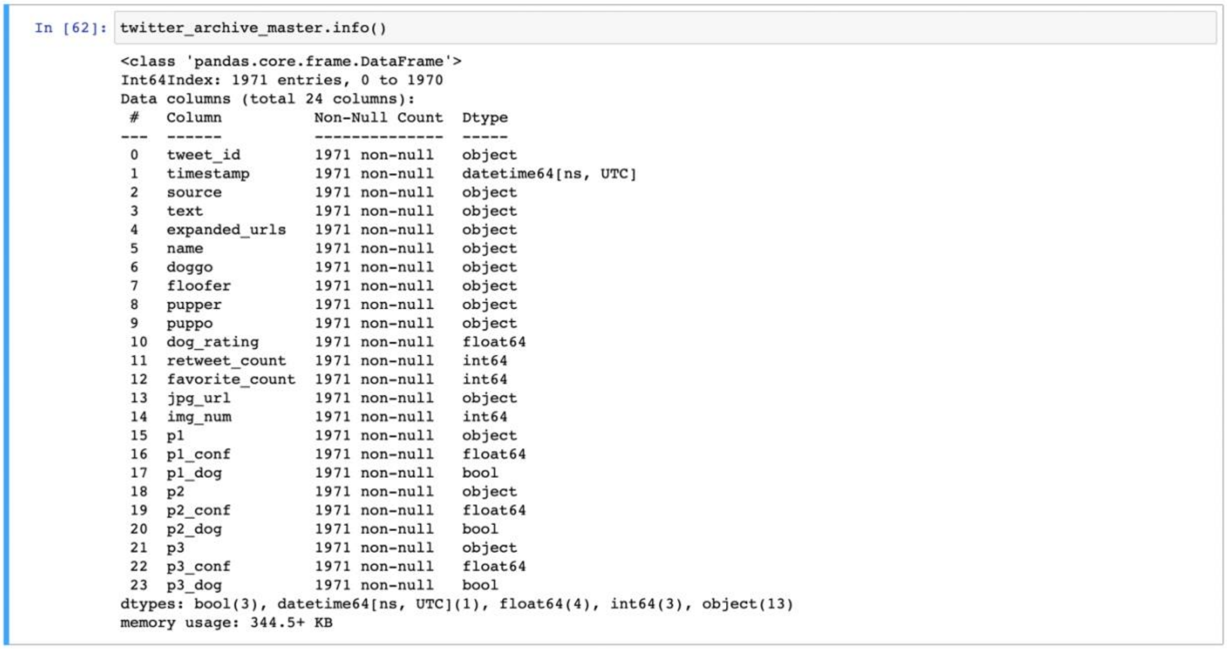


```
In [60]: image_predictions.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   tweet_id  2075 non-null   int64  
 1   jpg_url   2075 non-null   object 
 2   img_num   2075 non-null   int64  
 3   p1        2075 non-null   object 
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool   
 6   p2        2075 non-null   object 
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool   
 9   p3        2075 non-null   object 
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool   
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```



```
In [61]: tweet_clean.info()

<class 'pandas.core.frame.DataFrame'>
Index: 2354 entries, 0 to 2353
Data columns (total 3 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   tweet_id        2354 non-null   object
 1   retweet_count   2354 non-null   int64 
 2   favorite_count  2354 non-null   int64 
dtypes: int64(2), object(1)
memory usage: 73.6+ KB
```

Final twitter_archive_master dataset



```
In [62]: twitter_archive_master.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1971 entries, 0 to 1970
Data columns (total 24 columns):
 #   Column          Non-Null Count  Dtype              
---  ------          --------------  -----              
 0   tweet_id        1971 non-null   object             
 1   timestamp       1971 non-null   datetime64[ns, UTC]
 2   source          1971 non-null   object             
 3   text            1971 non-null   object             
 4   expanded_urls   1971 non-null   object             
 5   name            1971 non-null   object             
 6   doggo           1971 non-null   object             
 7   floofer         1971 non-null   object             
 8   pupper          1971 non-null   object             
 9   puppo           1971 non-null   object             
 10  dog_rating      1971 non-null   float64            
 11  retweet_count   1971 non-null   int64              
 12  favorite_count  1971 non-null   int64              
 13  jpg_url         1971 non-null   object             
 14  img_num         1971 non-null   int64              
 15  p1              1971 non-null   object             
 16  p1_conf         1971 non-null   float64            
 17  p1_dog          1971 non-null   bool               
 18  p2              1971 non-null   object             
 19  p2_conf         1971 non-null   float64            
 20  p2_dog          1971 non-null   bool               
 21  p3              1971 non-null   object             
 22  p3_conf         1971 non-null   float64            
 23  p3_dog          1971 non-null   bool               
dtypes: bool(3), datetime64[ns, UTC](1), float64(4), int64(3), object(13)
memory usage: 344.5+ KB
```