Reporting: wragle_report

The data wrangling project includes these tasks:

1. Gathering data

2. Assessing data

3. Cleaning data

4. Storing data

5. Analyzing, and visualizing data

1. Gathering data

The first task was to gather the data from three different datasets:

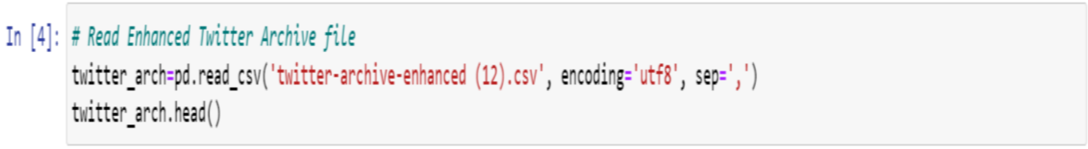
1.Enhanced Twitter Archive

2.Twitter API & json

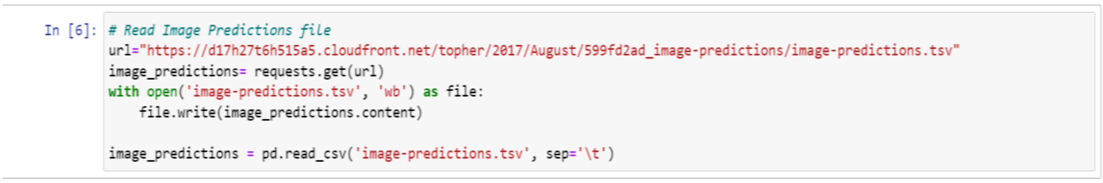
3.Tweet Image Predictions

The first, I downloaded the 'Enhanced Twitter Archive.csv' file manually and read it. Then, I download 'image-predictions.tsv' file programmatically by using requests library and the given url and create image predictions dataframe which consists tweet image predictions.

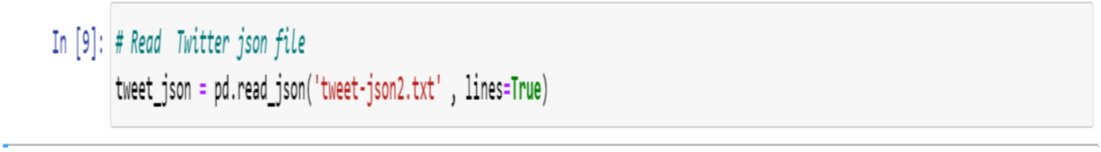
Finally, I download 'tweet-json2.txt' file manually, read it and create df_tweet data frame with tweet id, retweet count and favorite count columns and rename id to tweet_id.



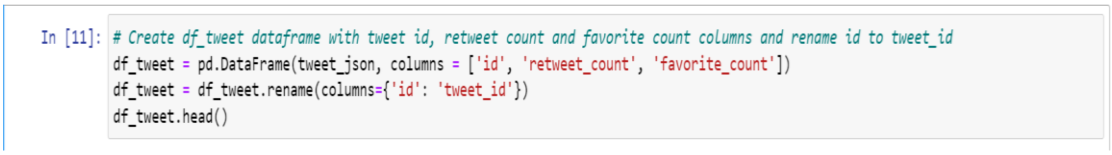
```
In [4]: # Read Enhanced Twitter Archive file
        twitter_arch=pd.read_csv('twitter-archive-enhanced (12).csv', encoding='utf8', sep=',')
        twitter_arch.head()
```



```
In [6]: # Read Image Predictions file
        url="https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv"
        image_predictions= requests.get(url)
        with open('image-predictions.tsv', 'wb') as file:
            file.write(image_predictions.content)

        image_predictions = pd.read_csv('image-predictions.tsv', sep='\t')
```



```
In [9]: # Read  Twitter json file
        tweet_json = pd.read_json('tweet-json2.txt' , lines=True)
```



```
In [11]: # Create df_tweet dataframe with tweet id, retweet count and favorite count columns and rename id to tweet_id
         df_tweet = pd.DataFrame(tweet_json, columns = ['id', 'retweet_count', 'favorite_count'])
         df_tweet = df_tweet.rename(columns={'id': 'tweet_id'})
         df_tweet.head()
```

2. Assessing data

After gathering data , I assessed the data programmatically by using pandas functions such as : info( ) , isnull() , duplicated() , value_counts() , unique() ,sample(), describe() and I found the quality and tidiness issues as the following :

Twitter archive table

Quality issues	Tidiness issues
1.Incorrect data types in the tweet_id , timestamp and retweeted_status_timestamp columns . *2.There are some rows have incorrect dog names like: "None”,"a" ,"an","Al","O".* *3.Must be merge rating_numerator and rating_denominator in one column (Rating) to become it easier to use it in analysis.* *4.There are some tweets after August 1st, 2017* *5.The dataset contains retweets.* *6.The source format must be reformatted in order for it to be readable.*	1.Must be merge the stages of dogs (doggo, floofer, pupper, puppo) in one column. 2.Need to be dropped some unnecessary columns.

Images predictions table

Quality issues	Tidiness issues
1.Incorrect data type in the tweet_id *2. There are mixed between capitalized and uncapitalized in dog breeds names.*	1. image_predictions_clean table must be merge to twitter_arch_clean table. 2.Need to be dropped some unnecessary columns.

Twitter API & json table

Quality issues	Tidiness issues
1.Incorrect data type in the tweet_id	1. df_tweet_clean table must be merge to twitter_arch_clean table. 2.Need to be dropped some unnecessary columns.

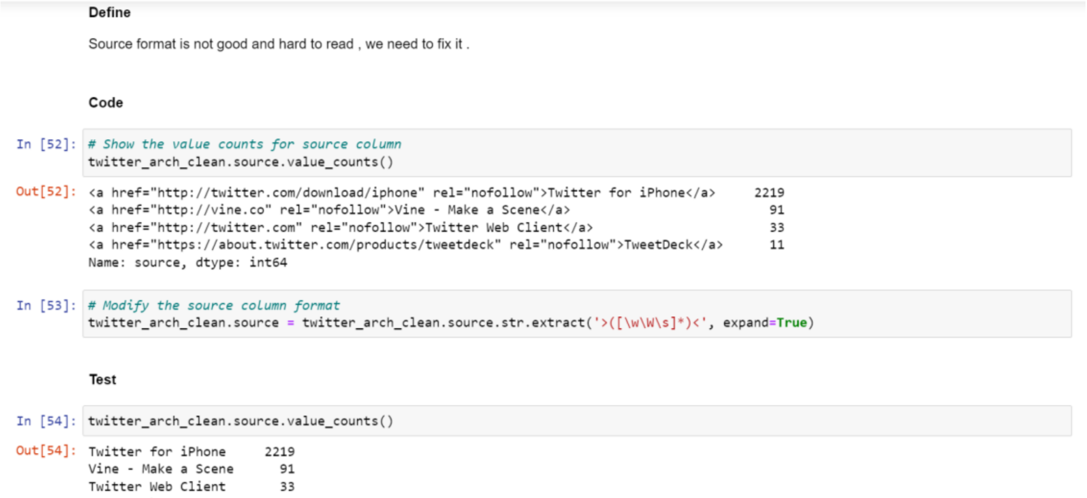
3. Cleaning data

First, must be created copy of three original data frames before clean it. Next, I cleaned data by documenting the define, code and test. The issues that I found through the assessment have been cleaned up using the following functions:

- astype( )   - drop( )   - extract ()   - info()   - head()

- capitalize()   - rename( )   -merge ()   -sum()   - match()

This is some examples for cleaning steps:



Define

Source format is not good and hard to read , we need to fix it .

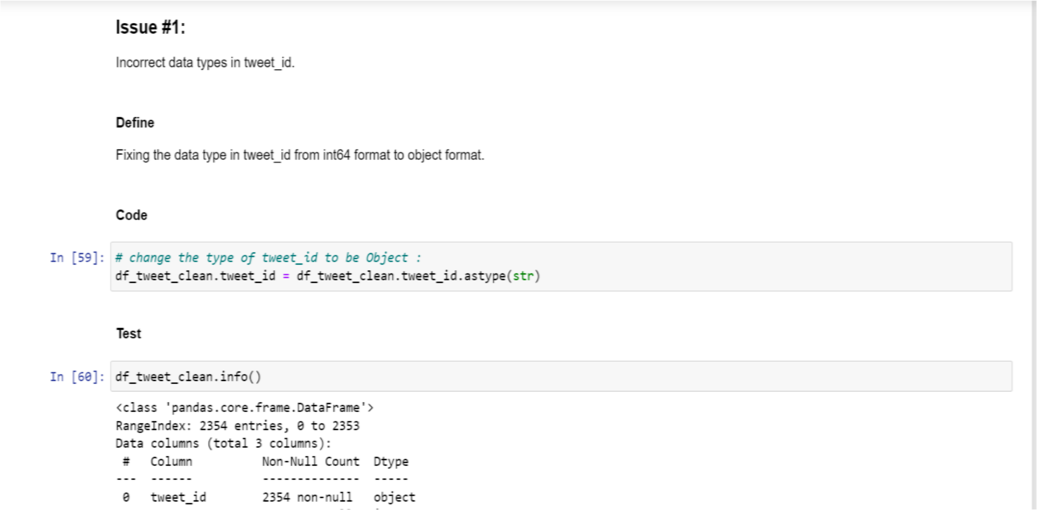
Code

```
In [52]: # Show the value counts for source column
         twitter_arch_clean.source.value_counts()

Out[52]: <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>     2219
         <a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>                          91
         <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>                       33
         <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>      11
         Name: source, dtype: int64
```

```
In [53]: # Modify the source column format
         twitter_arch_clean.source = twitter_arch_clean.source.str.extract('>([\w\W\s]*)<', expand=True)
```

Test

```
In [54]: twitter_arch_clean.source.value_counts()

Out[54]: Twitter for iPhone     2219
         Vine - Make a Scene      91
         Twitter Web Client       33
```



Issue #1:

Incorrect data types in tweet_id.

Define

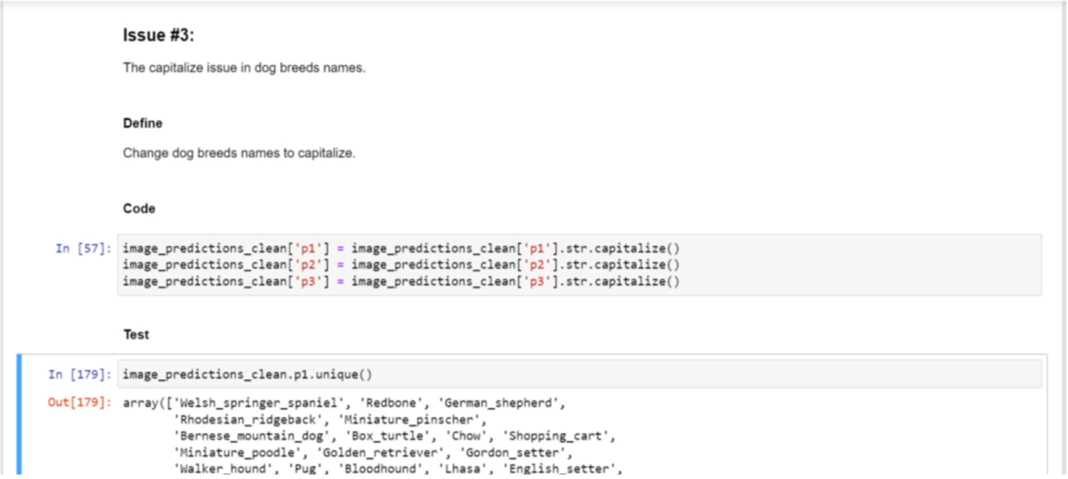
Fixing the data type in tweet_id from int64 format to object format.

Code

```
In [59]: # change the type of tweet_id to be Object :
         df_tweet_clean.tweet_id = df_tweet_clean.tweet_id.astype(str)
```

Test

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

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 2354 entries, 0 to 2353
         Data columns (total 3 columns):
          #   Column          Non-Null Count  Dtype
         ---  ------          --------------  -----
          0   tweet_id        2354 non-null   object
```



Issue #3:

The capitalize issue in dog breeds names.

Define

Change dog breeds names to capitalize.

Code

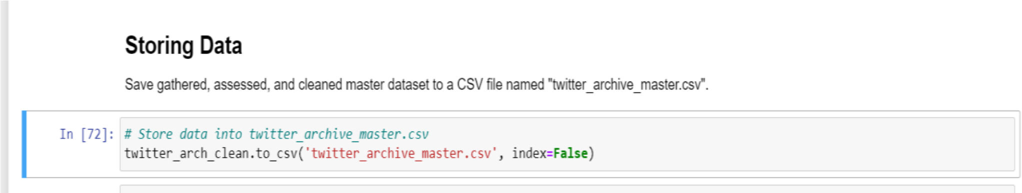
```
In [57]: image_predictions_clean['p1'] = image_predictions_clean['p1'].str.capitalize()
         image_predictions_clean['p2'] = image_predictions_clean['p2'].str.capitalize()
         image_predictions_clean['p3'] = image_predictions_clean['p3'].str.capitalize()
```

Test

```
In [179]: image_predictions_clean.p1.unique()

Out[179]: array(['Welsh_springer_spaniel', 'Redbone', 'German_shepherd',
                 'Rhodesian_ridgeback', 'Miniature_pinscher',
                 'Bernese_mountain_dog', 'Box_turtle', 'Chow', 'Shopping_cart',
                 'Miniature_poodle', 'Golden_retriever', 'Gordon_setter',
                 'Walker_hound', 'Pug', 'Bloodhound', 'Lhasa', 'English_setter',
```

4. Storing data

After cleaned data, now the dataset ready for analysis but before do it must be I store master table to twitter_archive_master.csv as shown in the photo :



Storing Data

Save gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".

```
In [72]: # Store data into twitter_archive_master.csv
         twitter_arch_clean.to_csv('twitter_archive_master.csv', index=False)
```