Data Wrangling Report

Project Objectives:

The main purpose of this project is to create a master dataset for WeRateDogs twitter account. The dataset will be used to bring some insights for them . Therefore, we will go through the following process:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

Step 1: Gathering Data

In this phase, we have to gather the needed data in order to create our master dataset. There were 3 sources that had been found:

1. Direct Source from WeRateDogs account: receiving the twitter archive dataset through email.
2. Finding a file on the internet: Downloading dogs breeds prediction file programmatically by using a Request library, the prediction have been applied on WeRateDogs' tweets.
3. Getting information from twitter API:

Step 2 and 3: Assessing and Cleaning Data

After gathering and reading the collected file into Pandas dataframe, we have been assessed the 3 datasets to find any quality and tidiness issues. So, the following is what have been identified and the proposed solutions:

Quality

Dataset	Issue	Solution
RatingDogs_df	there are 181 retweeted tweets, which they are duplicated tweets or from another users	drop all the retweeted tweets by using drop() function.
	there's 23 records with rating_denominator that is larger or smaller than 10, some of them is a true rating and others doesn't has the right rating extraction	since 99% of the rating dominator is 10 we will make it our standard value to extract the right rating by using regex and str.extract function, after that we will drop all the outliers since they will affect the analysis by using drop function¶
	name, dogge, floofer, pupper, puppo columns have none value as indicator of null	replace none values to null by using replace function
	most of the rows don't have a dog stage values, but i found out there's a number of values haven't been extracted	extract the values by using regex and str.extract function
	the name column has values that is not a name such as: a, an, the, and none as indicator of null	since the names has so many variety and is not a critical value in the analysis, we can convert them to null by using replace
DogsBreeds_df	There's a 66 duplicated rows	drop duplicated rows by using drop_duplicates function

	There's a lot of predictions that are not a dog breed and 324 rows don't have any dog breed prediction	drop all the rows that none of their predictions are dog breeds by using drop
All datasets	The datatype of ids and timestamp are not correct in all the datasets	fix the ids type by using astype(str) and timestamp types by using to_datetime

Tidiness

Dataset	Issue	Solution
RatingDogs_df	dogge, floofer, pupper, puppo columns are values not variable, so they should be in one column	getting the values from doggo, floofer, pupper, puppo columns and assign it in one column by using bfill() function
DogsBreeds_df	There's multipule breed prediction with verious confidence for each dog	Creating function to extract the right dog breed prediction for each image and apply it through all the rows
All datasets	the 3 datasets need to be mereged since all of them about the same thing which is the dog.	merge all the 3 datasets into one and choose only the necessary columns

Results

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2153 entries, 0 to 2152
Data columns (total 11 columns):
tweet_id              2153 non-null object
timestamp             2153 non-null datetime64[ns]
text                  2153 non-null object
rating_numerator      2153 non-null int64
rating_denominator    2153 non-null int64
name                  1422 non-null object
dog_stage             377 non-null object
jpg_url               1673 non-null object
dog_breed             1673 non-null object
favorite_count        1404 non-null float64
retweet_count         1404 non-null float64
dtypes: datetime64[ns](1), float64(2), int64(2), object(6)
memory usage: 201.8+ KB
```

	rating_numerator	rating_denominator	favorite_count	retweet_count
count	2153.000000	2153.0	1404.000000	1404.000000
mean	10.639573	10.0	8327.487179	2613.334758
std	2.250600	0.0	11188.622293	4066.836394
min	0.000000	10.0	52.000000	2.000000
25%	10.000000	10.0	1769.000000	578.500000
50%	11.000000	10.0	3904.000000	1311.000000
75%	12.000000	10.0	10348.750000	3117.750000
max	27.000000	10.0	107015.000000	56625.000000