Student name: Amani M. Ali

August 2020

Udacity Data Analysis Professional Track
Project (2): Data Wrangling project
We_Rate_Dogs

A) Gathering Data:

- ✓ The file 'twitter_archive_enhanced.csv' was downloaded manually.
- ✓ The image predictions file was downloaded programmatically using 'requests' library through the link provided on the classroom.
- ✓ Unfortunately, till the time of project submission, my request to twitter developer account was not yet reviewed and I didn't have access to the keys and tokens so I used tweet_json.txt file on the classroom instead to catch the project submission deadline.
- ✓ However, the code used to query the twitter API was revised, included and run.

B) Assessing Data:

The data was assessed visually and programmatically and the following issues were detected.

#	Issue Type	Issue
1	Tidiness	1. One variable in four columns(puppo, floofer,puppo and doggo.
2		2. All the three dataframes should be collected in one dataframe.
3	Quality	3. There is some retweets and tweets without pictures and duplicate rows.
4		4. There is some useless columns like retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp.
5		5. df1: Some rows has two dog stages at the same time.
6		6. df1: wrong values captured for numerators and denominators to be corrected manually.
7		7. df1: Decimal values were not correctly captured in the rating_numerator column.
8		8. df2: non-descriptive column names. ex:p1,
9		9. df1: timestamp is in string format,twitter_id integar not string
10		10. df1: wrong names for example: a, an, the??

C) Cleaning Data:

Every issue was appropriately addressed starting with the two structural issues followed by quality issues.

A new dataFrame was created by merging the three dataframes. The four columns representing the dog life stage were illustrated in one column. Rows containing multiple dog stages were concatenated by a comma.

Subsequently, content issues were addressed staring from removing duplicates and unnecessary data to correcting wrong values captured for rating denominator and rating numerator. Also, rating numerator extractor was modified to be able to capture decimal values from the text. Then, column names in the image prediction data were modified to more descriptive names.

Some variables were assigned to wrong data types. This issue was fixed by changing tweet_id from integer to string, timestamp from object to datetime and the stage from object to category. Also, rating numerator and rating denominator were previously changed to floats.

Then, the dog names column was cleaned. There were some wrong names extracted like ‘an, a, the’. Actually, all names that were not capitalized were not appropriate so they were all removed.