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.

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. Wrong values for denominator other than 10.
6		6. Very high inappropriate values in the rating_numerator column.
7		7. Non-descriptive column names. ex:p1,p2,p3.
8		8. df1: timestamp is in string format,twitter_id integar not string.
9		9. Wrong names extracted. For example: a, an, the??.
10		10. Some dog names does not start by a capital letter.

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.

Subsequently, content issues were addressed staring from removing duplicates and unnecessary data to correcting wrong rating denominator values and extreme rating numerator values. 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 dog_stage from object to category.

Then, the dog names column was cleaned for two issues. The first was wrong names extracted like ‘an, a, the’. The second was that some names were not capitalized.