Olalekan Ayinde

Introduction:

The tasks for this project are as follows: (1) Data Wrangling, which includes: (a) Gathering data (b) Assessing data (c) Cleaning data (2) Storing, analyzing and visualizing our wrangled data (3) Reporting on the following: (a) Our data wrangling effort (b) Our data analysis and visualization

The Dataset: The dataset contains tweet from WeRateDogs twitter archives. The archive has about 5,000 entries of their basic tweets. It also has 17 columns. For the purpose of this project, the dataset has been filtered to reflect tweets with ratings only. This now left us with 2,356 entries.
B. **Assessing Data:** Quality issues (at least 8) and tidiness issues (at least 2) were detected in the datasets. The issues are as follows:

Quality Issues¶

1. in_reply_to_status_id and in_reply_to_user_id have very scanty entries, we have to remove them. Moreover, they do not count much in our analysis.
2. There are incorrect dog name in under name column that need changes. They are: None, a, such, not, mad, an ,very, O, my, one, his, this, all, old, getting, the, by, officially, light, just, space, quite.
3. Remove many tweet under text column that are retweets. They start with RT or Retweet.
4. Remove all tweets that are reply to tweets.They start with @ or dot(.)
5. Timestamp column need to be changed from string to datetime for further processing.
6. Row 315 has 0 has the denominator. This need to be normalize to avoid dividing by 0. Its numerator of 960 is also too high a value for our dataset.
7. Outliers in rows 905 and 981 needs to be removed.
8. Sort out the required columns in tweet_json_df.

Tidiness Issues

Image Prediction

1. There is need to rename some of the columns in image_predictions_df to reflect what they do.
2. Removal of underscores in dog names. Rearrangement of columns and capitalization of dogs names.

The twitter-archive-enhanced-2.csv(we_rate_dogs_df), image-predictions (image_predictions_df) and tweet-json.txt (tweet_json_df) were assessed and the following actions carried out on them:
Check for duplicate rows: There are no duplicated rows for twitter-archive-enhanced-2.csv, image-predictions and tweet-json.txt.

C. **Cleaning Data**: Below are the cleaning of quality data issues as identified in the session above:

1. In_reply_to _status_id and in_reply_to_user_id have very scanty entries, we have to remove them. This is achieved by calling drop() function on the we_rate_dogs_df DataFrame.

2. There are incorrect dog name in name column that need changes. They are: None, a, such, not, mad, an ,very, O, my, one, his, this, all, old, getting, the, by, officially, light, just, space,quite. They are all in lowercase except None and O. There 109 entries that started with lower case and are incorrect dog names that needed to be dropped. Islower() function was used to sort names that are in lower case while drop() function was used to drop them.

'None' and 'O' were saved in the list of names to be removed. The function isin() was called on the list to removed them in name column.

3. Remove many tweet under text column that are retweets. They start with RT or Retweet. This was achieved by calling contains() function on 'RT' and 'Retweet' in we_rate_dogs_df DataFrame. 116 rows that contained 'RT' and 'Retweet' were removed.

4. Remove all tweets that are reply to tweets.They start with @ or dot(.). After calling startwith() on '@' and '.', there was no item that was returned, the text that begins with those words must have been removed while performing 3 above.

5. Timestamp column need to be changed from string to datetime for further processing. To_datetime() function was used to convert the content of timestamp column from string to datetime object.

6. Row 315 has 0 has the denominator. This need to be normalize to avoid dividing by 0. Its numerator of 960 is also too high a value for our dataset. All denominator that greater than or less than 10 were replaced with 10 by setting them to value 10.

7. The outlier of numerator value of 1776 in row 586 and everyone other rating above 50 has to be removed. All numerator that are greater than 11.0 (which is median) was replaced with it.

8. Sort out the required columns in tweet_json_df. These columns are id, retweet_count and favorite_count. This was achieved by creating a list and renaming the list name with the tweet_json_df.

The tidiness issues identified were also resolved as follows:

1. There is need to rename some of the columns in image_predictions_df to reflect what they do. This was achieved by using rename() function.

2. All the dog names need to be written properly, starting with upper case and removing underscores in image_predictions_df. Names in columns p1, p2 and p3 in image_predictions_df were capitalized. Using calling title() function. All underscores in names were removed using replace() function to replace underscore ("_") with space(" ").