WRANGLE REPORT

PROJECT TITLE: WE RATE DOGS TWITTER DATA

1.0 Introduction

This report contains detailed step-by-step process used in wrangling the twitter data set- We rate dogs. In this project 3 tables were utilized; these datasets were gathered differently. I sectioned this report into gathering, Assessing, Cleaning aspect, each section specifically explaining the process used.

2.0 Data Gathering

In this section, I directly gathered **all** three pieces of data for this project and loaded them in the notebook. However, the methods required to gather each data were different.

2.1 Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)

I directly gathered this data in that it was readily available in workspace, I used the notebook to read the csv data, then displayed the first five entry in the notebook.

2.2 Use the Requests library to download the tweet image prediction (image_predictions.tsv)

In gathering this data, i used request library to request the URL and stored it in a variable and then opened this file, using the text file. This data is a tab separated value, so i used the delimiter to open it up. I also included an if statement to prevent any error from popping up when opening again.

2.3 Use the Tweepy library to query additional data via the Twitter API (tweet_json.txt)

Additional data was gathered, this data included retweet count and favorite count. In gathering this I was supposed to query twitter API, but in doing this, I didn't get a response to my request to get a twitter developer account. So, I settled for gathering the data manually, in which I did creating an empty list, and then opening the JSON file as txt file, filling the empty list with the info from the JSON file. Then I went ahead to display the first five data in the notebook using the – (.head()).

3.0 Assessing Data

In this section, after gathering the required data, I assessed the data to detected possible quality and tidiness issue. I detected and documented at least **eight (8) quality issues and two (2) tidiness issue**. Using **both** visual assessment programmatic assessments to assess the data.

I paid rapt attention to the following key points when accessing the data.

- You only want original ratings (no retweets) that have images.
- The requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned.
- You do not need to gather the tweets beyond August 1st, 2017. In assessing this table, I did it from table to table while documenting my observation for each. I assessed the tables visually and programmatically and then sectioned each observation into Quality and Tidiness issues.

3.1 Assessing Twitter Archive Table

After visually and programmatically assessing the above table, this were the result, deduced.

- The following columns - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp have null values. And these columns don't really contribute to the analysis process. Let's assess these columns to understand why they are NAN values. Using (.info()) and assessing each columns.
- For the retweeted_status_timestamp column, we can't get back these values. Thus, we don't need them that much since we have a complete timestamp column. The other columns above have incomplete values, and can't be addressed using imputing, so I'll rather take them off. Used the dtype function to check for data type

Quality
Tweet_id has invalid datatype. timestamp column has invalid data type expanded_urls column has missing values
The value for the source column is inaccurate
The values for the rating_numerator column are inconsistent
Name column has invalid value (a) and None.

Tidiness
The twitter_archive table has an issue of tidiness, where each observation should form a row –
The dog type, doggo,floofer, pupper, puppo are variables and should form a column.

3.2 Assessing Image Prediction table.

In this table, I assessed visually and programmatically. and the results are portrayed below.

- Visual Assessment: image_prediction[image_prediction['p3_dog'] != True]

Quality
Inconsistent column names for p1, p2, p3, p_conf, p_dog columns p1, p2, p3 values are separated by hyphen character Values for p1,p2,p3_conf are in proportions instead of percentage.
Columns that has all its p_dog values to be False should be eliminated.
The columns p1, p2, p3, p3_conf, p2_conf, p1_conf, should be streamlined to one column, depicting the correct name of the picture.

- Programmatic Assessment

Quality

The column tweet_id is an integer data type p1_conf, p2_conf, p3_conf are columns that should be change to percentage.

Columns with p names, should be streamlined to have one column.

3.3 Assessing retweet_fav_count table

In this table there weren't any quality or tidiness issue.

Visual Assessment: retweet_fav_count.tail() Pretty good to go visually.

Programmatic Assessment: retweet_fav_count.dtypes Pretty good to go.

4.0 Cleaning Data

In this section, I cleaned **all** of the quality and tidiness issues I documented while assessing the tables. However, I ensured that I made a copy of the original data before cleaning. Cleaning included merging individual pieces of data according to the rules of tidy data
I ensured that the result of cleaning is a high-quality and tidy master pandas Data Frame. I also ensured that the cleaning dimension I maintained allowed me to first clean missing data issues, then proceeded to cleaning Tidiness issue then finally went ahead to clean the remaining quality issues

4.1 Cleaning dimension

- Made copies of original pieces of data. I did this to prevent missing all the documented information.

- Cleaned Missing values first.
  Missing Values: Unwanted Columns to be eliminated Missing values #2: Expanded Url has missing values
- Cleaned Tidiness issues.
  Tidiness: Image_prediction table: Each observation should form a row Drop unwanted columns - img_num, p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog, p3_dog, Dog_category Eliminated rows that don't have dog breed Tidiness: Develop a column in which each variable (Doggo, floffer, pupper, puppo) makes a column. Tidiness: Twitter_archive and image prediction table should be merged together.
- Finally, cleaned the remaining quality issues.
  We now have 2 tables - twitter_archive_image and retweet_fav_count_clean Quality: Duplicate values in the twitter_archive_image table Quality: Invalid datatype Quality: Inaccurate value for source column Quality: Fixing Text column Quality: Inconsistent values for rating denominator and rating numerator Dog breed values are separated by columns

5.0 Storing Data

In this section, I saved the gathered, assessed, and cleaned twitter_archive_image, retweet_fav_count dataset to a CSV file named "twitter_archive_master.csv", "retweet_count.csv" respectively.