Project 2

Wrangle Report

ALX Nanodegree Data Analyst Program

Elaborated by: Oumayma Jribi

Table of content

1. Introduction:

This project aims to wrangle, analyze, and visualize the WeRateDogs dataset, which is the tweet archive of Twitter user @dog_rates. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

2. Data Gathering:

The data for this project consists of 3 different datasets that were obtained as following:

- **Enhanced Twitter Archive**: obtained from the twitter_archive_enhanced.csv file provided by Udacity which was manually downloaded.
- **Image Predictions:** which is a file that contains a dog breed prediction for each dog in a certain tweet. 3 predictions are given for each dog with their corresponding confidence level and whether that prediction is true or not. This file was hosted on Udacity's servers and was downloaded programmatically using the Requests library and he given URL.
- **Twitter API and JSON:** twitter API was used to query additional data beyond the data included in the WeRateDogs Twitter archive. This additional data will include retweet count and favorite count. The data gathered was written to a JSON file tweet_json.txt to be used later. There was also the option to access this data without a twitter account by downloading the tweet_json.txt directly as given by Udacity.

All along this project, the data frames used was named as following:

- **dog_archive** : containing the enhanced twitter archive data
- **image_predictions** : containing the dog breed predictions data
- **tweet_data** : containing the additional retweet count and favorite count gathered using twitter API and written to the tweet_json.twt file

3. Assessing Data:

Once the 3 tables were ready to use, the data was assessed using:

- Visual assessments: by printing the dataframes and trying to go over them and identify issues to be cleaned at different phases of the assessment step
- Programmatic assessments: using functions and methods (e.g info(), describe(), sample, duplicated, value_counts…) that reveal issues with the data's quality and tidiness

Issues identified thanks to the assessment phase were noted in a separate section of the jupyter notebook and divided to **Quality Issues** and **Tidiness Issues,** as well as divided by the tables to which each issue belongs.

4. Cleaning Data:

The define, code, and test methodology was used to structure the cleaning step.

Before starting cleaning the datasets, a copy of each was made. Cleaning steps later were done to these copies, initial datasets were preserved.

The issues I identified and cleaned are the following:

- **Quality issues**

dog_archive

1. Wrong values in the 'name' column in dog_archive table ('a', 'the', 'an', 'very', 'quite', 'just, 'one' ....) should be removed and replaced with NaN
2. Timestamp column should be of type datetime instead of object
3. Retweets (tweets with a value in retweeted_status_id) should be dropped
4. "None" values in all columns should be replaced by NaN to indicate values that are not available
5. Denominators different from 10 should be changed to 10
6. Tweets with no expanded urls should be dropped
7. Reformat the 'source' column

image_predictions

8. Remove underscores from dog type predictions and make all dog names proper nouns (start with capital letters)

- **Tidiness issues**

1. Tweet_data should be part of the 'dog_archive' table
2. image_predictions should be part of the 'dog_archive' table (by creating a new column 'dog_breed' in dog_archive containing the probable breed for each dog)
3. doggo, floofer, pupper, and puppo columns in dog_archive should be melted to one 'dog_stage' column
4. retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns can be dropped from dog_archive since all retweets will be deleted

By the end of the cleaning step, the main dataset was the dog_archive_clean dataset, to which many columns from the other 2 datasets were added.

5. Storing Data:

The cleaned dog_archive_clean dataset was stored to **twitter_archive_master.csv** file using **.to_csv** function.

6. Data Visualization and Insights:

The final cleaned dataset was used to derive insights and make visualizations. These latter will be presented in the act report file.

7. Project Challenges and Limitations:

Working with twitter API was the biggest challenge for me during this project. I did all the steps needed to make it work but I was given essential access which didn't cover some of the main functions used to gather the data needed for this project. I kept getting the "Forbidden" error and that I needed to have elevated access. Fortunately though the tweet_json.txt file was provided to make this task easier.