We Rate Dogs on Twitter

Programm: Udacity Data Analyst Nanodegree (DAND)
Student: Laman Mammadli
Project: Data Wrangling

Introduction

I have completed this project as an Udacity student who takes Data Analyst Nondegree. Below, I will describe the steps I took to wrangle the data.

Gathering Data

For this project, Udacity provided these files:

- twitter_archive_enhanced.csv
- Image_predictions.tsv
- tweet_json.txt

I downloaded the first two files from the Udacity Project Details page. Additionally, I need the “WeRateDogs” Twitter archive from the Twitter API. As recommended, I applied for the Twitter developer account to get the data. Unfortunately, I got the rejection. *In the end of this report, you can see the screenshot of the email I got from Twitter. Fortunately, Udacity had been provided the tweet_json.txt file in the supporting materials. And I used it for the project. After downloading all three files, I uploaded them to the Jupyter Notebook.

Assessing Data

After uploading the files to the Jupyter Notebook, I imported essential libraries (pandas, numpy, seaborn, matplotlib) for the wrangling of the data. Then by means of pandas I read the three data files and loaded them to these tables; “twitter”, “image”, “tweet”. I assessed the tables in sequence.

“twitter” table

Quality Issues

Initially there were 2356 rows, 17 columns. While assessing, I identified quality and tidiness issues:
- Wrong data types: timestamp, retweeted_status_timestamp are object, not datetime64
- Missing data (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)
- Missing names, nulls represented as 'a' instead of real ones in the name column
- Missing information, nulls represented as 'none' in columns such as doggo, floofer, pupper, puppo.

- There are 23 rows which contains rating_denominators is not equal 10.

Tidiness Issues

- tweet_id column in `twitter` table duplicated in `image`
- source, in_reply_to_status_id and in_reply_to_user_id columns in `twitter` table duplicated in `tweet`
- there are unneeded columns
- date and time are under one variable - timestamps

“image” table

Then, I started assessing this table. Initially there were 2075 rows and 12 columns. Below, you can see the findings on quality and tidiness issues:

Quality issues

- Incomprehensible column names (p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog, p3_dog)
- There are several names assigned to the dogs (p1, p2, p3) which are not dog names, such as hen(p3), toilet_tissue(p3), computer(p1), paper_towel(p2), and so on
- There is no standard form of writing the predictions in p1, p2, p3. Lower or upper cases, underlines and so on

Tidiness issues

- tweet_id in `twitter` and `image` tables are matching with the id column in `tweet` table and can be merged after renaming the id column to tweet_id columns in `tweet` table

“tweet” table

Lastly, I assessed the “tweet” table and found out quality and tidiness issues:

Quality issues

- Missing values (extended_entities, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, geo, coordinates, place, contributors, possibly_sensitive, possibly_sensitive_appealable, retweeted_status, quoted_status_id, quoted_status_id_str, quoted_status)

Tidiness Issues

- id and full_text columns in `tweet` table are matching with the tweet_id and text columns in `twitter` table
- there are unneeded columns
- date and time are under one variable - created_at

Cleaning Data

After identifying the possible number of quality and tidiness issues in these three table, I started the cleaning and washing. Firstly, I dealt with the all quality issues I identified in three tables, and after that I looked at the tidiness issues.

In the `twitter` table, I converted the timestamp column data type from object to datetime64. There were null values in several columns, which I dropped them. Under the name column, there were only several names for the dogs, and others were unrelated text such as “a”, “none”. I dropped it. Then I dropped the stage name columns, which contained “None” text, which is not useful. Then I find the indices of the rows which contains rating_denominators higher or lower than 10, which should not. I collected them in a list. And then dropped those rows.

In the `image` table, I renamed the columns which are not descriptive. Then in the predictions I found that there were names which are not dog names, I found those rows and checked the photos. Several of the photos are not dog photos. I dropped those rows. Moreover, in those columns there were not a standard writing of dog names. I make all of them lower case and replace the underscores with the white space.

In the `tweet` table, I dropped several columns which have null values. Moreover, renamed the id column as tweet_id.

After completing the cleaning on quality issues, I merged all three table on tweet_id column. Then I assessed the new table (twitter_clean) and dropped one of the doubled columns, such as source_x and source_y, timestamp and created_at, text and full_text. There were also other columns which were not useful for my project. I dropped them too. Then, from the timestamp column, I extracted the date and called the column tweet_date. Dropped the timestamp.
After all these cleaning, I saved the table into twitter_archive_master.csv file. Then read it again, and loaded the data into twitter_master table. Then I converted the tweet_date data type from object to datetime64.

The new table twitter_master is ready for the analysis phase.

*Twitter rejection for Developer Account