Project Four: Wrangle and Analyze We Rate Dogs

Date : March 5, 2021
Student : Rodrigo Contreras Vielma
Email : contrerasvielma@gmail.com

Wrangle Report

My wrangle efforts were (Using Jupiter Notebook):

1) Gather Data

- I read the main dataset from csv file "twitter-archive-enhanced.csv" provided by Udacity (using pandas library).
- I Downloaded the second dataset from URL provided by Udacity, using get: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv (File image_predictions.tsv, using requests library).
- I created my twitter account and after subscribe to twitter developer account to use Twitter API.
- I connected with Twitter API to get file "tweet_json.txt". I had some problems with no access to some tweets. Finally, to avoid any kind of trouble with the data, I used the original file provided by Udacity (tweet-json_original_Udacity.txt). Was a great experience to use the tweeter API.
- I read the datasets using Visual studio code, as text editor, to see the data rows.

2) Assess Data

- I assess different dataset using python command: head, tail, sample, info, duplicated, sum, value_counts, lambda functions, isnull, describe.
- I wrote the following first notes about the data: Convert datatypes, drop rows, drop columns, strange patterns with names, anormal values from numerators and denominators, columns not properly designed, replace control characters to export to csv and database Sql Lite, lower and capital case, duplicated pictures, rename columns, consolidate data from many columns, merge all datasets.

3) Cleaning Data

- I made a copy of the different datasets.
- I applied the method: Define, Code, Test in each process of cleaning and tidy data.
- The step to cleaning data were:

Quality Issues

- *df_tae_clean (twitter archive enhanced):*
  - *Convert Tweet_id to string.*
  - *Convert timestamp to datetime.*
  - Drops tweet with retweets. One patterns is when the text start with the string"RT @".
  - *Drop rows that don't have expanded_urls.*
  - *Drops colums `in_reply_to_status_id`, `in_reply_to_user_id`, `source`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp` (not necessary to our analysis).*
  - *We have many record with a trouble with names "a", "the", "an", "officially". The pattern is where the first letter begin a lower case.*
  - *Clean "\n" in column `text` to export in one line.*
  - *Correct `rating_numerator` because it was not reflecting the numerators with decimal point well.*
  - *Consolidate rating in one column `rating_score`.*
  - *Drop columns `rating_numerator`, `rating_denominator`.*
  - *Change name of column text to avoid error and confusion with datatype text of SqlLite3.*

- *df_ip_clean (image predictions):*
  - *Convert Tweet_id to string.*
  - *We appreciate that the columns p1, p2 and p3 are not consistent with capital case or lowercase. So I am going to convert all to lowercase.*

  - *Delete rows that correspond to retweets.*
  - *Create two new columns with breed of dog and level of confidence. After clean columns. Adjust datatype to float (confidence), after round with 4 decimals.*

- *df_tj_clean (Tweets Json):*
  - *It's has a big structure, we only get the columns id, favorite_count, retweet_count.*
  - *Column id will be renamed to tweet_id and convert to string.*

Tidiness Issues

- *df_tae_clean (twitter archive enhanced):*
  - *We need to consolidate in a one column the four columns (doggo, floofer, pupper, puppo). We applied a new function to concatenate some cases with rows that have values combined.*
  - *Merge all dataframes.*

The previous steps were an iterative process and I was a very good experience. I researched and studied a lot of tutorials to reach the goal of this project (I invested a lot of time in this and although I am behind in the project I learned too much). I left the reference (URL) en each sentence where I found the method of programming that let me complete the process.

4) Data Storage:

I saved:

- The consolidates the master dataframe in csv format using pandas (twitter_archive_master.csv)
- Create a database SQL Lite 3.

5) Data Visualization:

- Finally, I created a lot of charts with insight about de data:
    - Graph bar by King of dog, Histogram Name of dog, breed of dog
    - Correlation between retweet_count and favorite_count (To close).
    - Correlation between rating_score and retweet_count (Bad correlation).