Data Analyst Nanodegree | Udacity

Project 4: WeRateDogs Twitter Data Analysis

Completed by Alex Maksimov

In order to wrangle the data I had to gather, assess and clean the data to prepare it for the analysis and find new insights.

Gathering

First, I needed to gather the data. The data I needed consisted of 3 parts:

1. twitter_archive_enhanced.csv - a flat file that contains the data about dog tweets and is the main data file in our analysis
2. tweet_json.txt - a json file that was composed via connecting to twitter API. This file contains the data with retweet counts and favorite counts.NOTE: I requested access to twitter API by submitting the form but have never received anything back from them and I had to use the file provided by Udacity.
3. image-predictions.tsv was downloaded from the link provided by udacity by means or requests library in Python. This file contains dog breed names generated via machine learning algorithm.

After gathering the data I saved tweet_json.txt file to a .csv to be able to process it in Google Sheets later.

Accessing

I used Google Sheets for the visual assessment and built-in functions for the programmatic assessment. In Google Sheets I opened each of the files and looked for errors in data, e.g. I found that the rating denominator is 00 with Google Sheets. After dealing with Google Sheets I assessed the data programmatically.

During the assessment I found the following issues:

Quality:

twitter_archive_enhanced.csv (‘archive’):

- mistakes in 'name’ column.
- 'timestamp' column has `+0000` substring along with datetime values.

- some ints in 'rating_numerator' and 'rating_denominator' columns are 3-digit ones.
- a record with ‘tweet_id’ = 835246439529840640 has incorrect rating - 960/00.
- hyperlinks are present in the 'text' column.
- 'tweet_id' is an int not a string as it should be since the numbers in this column aren’t meant to be used in calculations.
- 'timestamp' is a string not a datetime as it should be.
- values in the 'source' column are surrounded by html tags which makes them inconvenient to analyze.
- some dog tweets have 2 dog stages.
- columns indicating dog stages are objects not categories.
- 'source' column is an object not a category.

tweet_json.txt(‘counts’):

- 'id' is an int not a string.
- 'favorite_count' and 'retweet counts' are floats even though this column contains only integers.

image-predictions.tsv(‘image’):

- 'tweet_id' is an int not a string.
- some values in the 'p1_conf' column are FALSE and don't represent the breed of a dog which means these values won’t be needed in the analysis.
- inconsistent dog breed names (some are in lower case and some capitalized).
- dog breed names are separated by underscore.
- 'p1' column should be renamed as 'breed'.
- columns indicating dog breeds is an object not a category.

Tidiness:

- values in 'doggo', 'floofer', 'pupper', 'puppo' columns in ‘archive’ data frame belong to the same variable.
- in all 3 dataframes there are columns that aren’t necessary for analysis.
- all 3 data frames are the parts of the same observational unit (tweets with dogs).
- 'rating_numerator' and 'rating_denominator' values in ‘archive’ data frame should be in one column since they represent the same variable.

Cleaning

Before proceeding to data cleaning I created copies of each data frame so that I can revert the changes should I make a big mistake. Then I did the following:

- Dropped all the columns that I won't need in the analysis. Some of these columns (like 'in_reply_to_status_id', 'expanded_urls',etc.) seemed meaningless and others (like 'name') contained a lot of errors and couldn't be used in the analys.
- Pivoted 'doggo', 'floofer', 'pupper', 'puppo' columns in 'archive' data frame and created one column populated with values from those 4 columns.
- Removed the '+0000' substring from values of the 'timestamp' column in 'archive' data frame since it doesn't carry any meaning and hinders from converting this field to datetime type.
- Corrected the value in 'twitter_id' = 835246439529840640 in 'archive' data frame since it's obviously incorrect and can cause the zero division error in calculations.
- Removed hyperlinks from the 'text' column and html tags from the 'source' column in the 'archive' data frame since they didn't provide any meaningful information.
- Removed html tags from the 'source' column in the 'archive' data so that the values can be analyzed.
- Renamed a 'p1' column (indicating the most accurate prediction of dog breeds) to 'breed' in 'image' data frame, capitalized the breed names and removed the underscores between parts of breed names.
- Filtered the values in the 'image' data frame so that only 'p1_dog' = True left.
- Merged all the tables into 1 table, converted the data types in the fields accordingly and created a new column from the division of 'rating_numerator' column by 'rating_denominator' column.
- I didn't correct 3-digit values in 'rating_numerator' and 'rating_denominator' columns since they were supposed to be divided and the ratio won't differ a lot from 2-digit ratings.

After doing all of these I considered the data ready for the further analysis and saved it in the twitter_archive_master.csv file.