Udacity Data Analyst Nanodegree

Vasileios Garyfallos - April 2020

1. Introduction

In this document, I will briefly overview my approach to the data wrangling process.

Before wrangling the data, I imported all relevant libraries, namely:

Requests
Numpy
Pandas
JSON
Seaborn
RE
Tweepy
Matplotlib

2. Data Wrangling

2.1 Gather

The data itself could be gathered from three sources:

a) a .csv file, provided by Udacity — twitter-archive-enhanced.csv
b) a hyperlink — I requested the file's URL (via Python Requests library) and wrote the content of the response to a .tsv file
c) Twitter API — I created a Twitter developer account, in order to gather the data from the Twitter API. Then pulled the data using Python Tweepy library. I wrote the data in a new file called "tweet_json.txt".

I have since removed my credentials and to re-run the notebook, another user’s credentials will need to be entered.

Unfortunately, not all of the tweets were accessible. Any tweet ID which raised an error was put into its own dataframe. I examined this dataframe and found that in 24 of the 25 cases, no tweet (referred to by Twitter as a “status”) was found matching the tweet ID. It’s possible that those tweets were deleted or the tweet ID was incorrectly recorded. In the last case, I was not authorized to see the tweet – perhaps because someone set their account to private. Either way, I was unable to work around these errors and unfortunately the 25 tweet IDs were later deleted from the final datasets.

2.2 Assess

Once all the data was collected, I used a combination of visual and programmatic tools to assess the dataframes. There was a lot of missing data in the columns about the reply and the retweeted status. Since we only wanted original posts with images I had to drop them.

Quality assessment:

df_twitter table

- the datatype of the id - columns is integer and should be str
- the datatype of the timestamp - column is object and should be datetime
- some of the dogs are not classified as one of "doggo", "floofer", "pupper" or "puppo" and contain all "None" instead
- some of the dog names are not correct (None, an, by, a, ...)
- contains retweets
- some of the ratings are not correctly extracted (mostly if there are >1 entries with the pattern "(\d+(.\d+)?/\d+(.\d+)?)"
- also transforming the ratings to integer created some mistakes (there are also floats)
- the source column contains html code

df_predict table

- the datatype of the id - columns is integer and should be str
- contains retweets (duplicated rows in column jpg_url)
- there are pictures in this table that are not dogs
- the predictions are sometimes uppercase, sometimes lowercase
- also there is a "_" instead of a whitespace in the predictions

df_api table¶

- the datatype of the id - columns is integer and should be str

Tidiness assessment:

***df_twitter* table**

- the columns doggo, floofer,pupper and puppo are not easy to analyze and should be in one column

***df_predict* table**

- the prediction and confidence columns should be reduced to two columns - one for the prediction with the highest confidence (dog)

***df_api* table**

- display_text_range contains 2 variables

***all* tables**

- All three tables share the column tweet_id and should be merged together.

2.3 Clean

Cleaning steps:

1. Merged the tables together
2. Dropped the replies, retweets and the corresponding columns and also dropped the tweets without an image or with images which don't display doggos
3. Cleaned the datatypes of the columns
4. Cleaned the wrong numerators - the floats on the one hand (replacement), the ones with multiple occurence of the pattern on the other (drop)
5. Extracted the source from html code
6. Splitted the text range into two separate columns
7. Removed the "None" out of the doggo, floofer, pupper and puppo column and merge them into one column
8. Removed the wrong names of name column
9. Reduced the prediction columns into two - breed and conf
10. Cleaned the new breed column by replacing the "_" with a whitespace and make them all lowercase

3. Analysis

The remainder of my notebook is dedicated to reshaping and filtering my dataset for the purposes of analysis. The analysis is based on 3 questions/metrics:

- Based on the predicted, most likely dog breed: Which breed gets retweeted and favorited the most overall?
- How did the account develop (speaking about number of tweets, retweets, favorites, image number and length of the tweets)?
- Is there a pattern visible in the timing of the tweets?

The questions are answered separately and are accompanied by various plots.