Wrangle Report

Author: Jaroslav Tran
Date: 06/03/2020

Data Sources and Gathering

There are 3 sources that have been used for this data analysis:

1) WeRateDogs dataset that has been exclusively given to Udacity for this project in the form of a csv file. The name of that dataset is twitter_archive. It contains over 5000 tweets as of 1st of August 2017. I downloaded this dataset and accessed it in the jupyter notebook.
2) The second dataset was image_tweets which is dataset of predictions using convolutional neural network on the images of the dogs to classify them into their breeds. I programmatically downloaded this dataset using request Python library and downloaded it as a tsv file
3) Lastly, using tweet IDs from the twitter_archive (our first dataset), I queried the Twitter API to get each tweet´s data in JSON format. For that I used a Tweepy library. I stored each tweet´s entire data that were later further processed

Data Evaluation

Once all the datasets were available, I investigated the dataset using both visual and programmatic assessment.

1) Visual Assessment

- I printed the heads (first five rows) and tails (last five rows) of each dataset to observe clean discrepancies in data quality or tidiness.

2) Programmatic Assessment

- I used a variety of pandas methods such as:
    - .info()
    - .value_counts()
    - .isnull().sum()
    - .duplicated()
    - .groupby()

Afterwards I have written down and categorized issues into data quality or data tidiness depending on the character of the problem at hand.

Cleaning Process

Before the cleaning, I made a copies of the original data frames to have a safeguard against a potential mistakes.

The cleaning process itself followed a standard data cleaning process of defining the issue, coding a resolution to it and testing the code to see if it was resolved.

The data quality issues that I have resolved were:

twitter_archive

- It seems that the twitter_archive dataset contains not only the original tweets but also retweets with images. Since we want to analyse only original posts and not retweets, we will have to get rid of the retweets.
- Incorrect data types: in_reply_to_status_id and in_reply_to_user_id should be objects(string) not floats. Timestamp should be a date not an object.
- I suspect that the numerators and denominators given might not match the ones in the text. (The ones that are not multiples of 10)
- The rating denominator is an object but should be an integer if we want to do some calculations with it
- Some of the dog names are misspelled or mislabeled so we will have to either convert them into None or read them from the Text column

image_predictions

- I suspect that there might be consistency issues with the dog breed names (p1, p2, p3)
- All the headers with the names p1, p2, p3 are not very informative so we will be renaming them do prediction1, prediction2,..

tweet_json

- There are some retweet data in this dataframe so I will have to get rid of them as well
- Tweed id needs to be converted into integer to be able to merge all the datasets together

The data tidines issues that I have resolved were:

twitter_archive

- There are four stages of a dog rating (doggo, floofer, pupper, puppo) that could fit into one column
- The twitter_archive, json_tweets and image prediction datasets should be merged into one as they are giving information about the same twitter id accounts to adhere to the tidiness principles:
    - Each observation forms a col
    - Each observation forms a row
    - Each observational unit forms a table

Conclusion

Python and associated packages (pandas, numpy, tweepy) were used to gather, assess and wrangle the data.

The most challenging parts of the project were:

1) Scraping the data from Twitter API using Tweepy was challenging and time consuming since every query took over 20 minutes.
2) Finding the issues. Some of them were obvious but quite of them would be easy to overlook (and have been overlooked).
3) Melting the dog stages into one column instead of keeping four columns for each dog stage in the twitter archive dataset,