DATA ANALYST NANODEGREE

WeRateDogs Twitter Analysis

INTERNAL REPORT

ßdog_rates · Apr 16
m told she enjoys car rides and smells good. 12/10 would
she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Introduction

This report summarises my wrangling efforts on the WeRateDogs Twitter Analysis. I separated this document into the three steps in data wrangling; gathering, assessing and cleaning data.

Gathering

First I had to gather data from three sources. The first one provided was an archive of the account given to Udacity which I manually downloaded. The second was a dataset containing images from each tweet with a neural network's predictions on the breeds, the confidence interval and whether or not the prediction was an actual breed. This dataset was hosted on Udacity's servers and had to be programmatically downloaded with Python's request library. Due to the archive not containing vital information on the number of retweets and likes for each originally rated tweet, I had to scrape the data from Twitter. To do this, I applied for a Twitter developer's account which was granted to me after 3 tries. Then I was able to generate secrets and tokens that enabled me to scrape the tweet into a .txt file. Afterwards, I converted the file into a .csv file.

Assessing

In accordance with what I learnt, I assessed the data visually and programmatically. For my visual assessment, I used both Microsoft Excel and calls to the variables holding the dataframes. Then I programmatically assessed using various methods. With this, I was able to spot 11 quality issues and 2 tidiness issues listed below:

Quality issues

In the manual_df table:

1. tweet_id is an integer not a string
2. many missing records for in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns
3. timestamp is an object not timestamp
4. name column seems to have wrong names and stop words instead of names i.e 'a', 'old', 'one', 'the'

5. (Given in classroom) - some records are retweets, we only want original ratings with images
6. There are many None values in puppo, pupper, floofer and doggo columns but no null values when checked with .info()

In the image_pred table:

1. tweet_id is an integer not a string
2. inconsistent casing in p1, p2, p3 columns
3. p1, p2, p3 column names are not well descriptive

And lastly, the subtweet_data table:

1. id is an integer not a string
2. id is represented as tweet_id in other tables

Tidiness issues

1. One variable in four columns (puppo, pupper, floofer and doggo) in manual_df table
2. The archived twitter dataset and the newly scraped twitter dataset tables should be joined

Cleaning

In order to produce high-quality and tidy datasets, I used a variety of methods from pandas and numpy to clean the datasets and then stored them in .csv files for future reference.

Conclusion

For full technical details of my performance on data wrangling, refer to the 'wrangle_act.ipynb' file.