Udacity Data Analyst Nanodegree Program

Wrangling process

WeRateDogs

Introduction

WeRateDogs is a twitter account that rates people's dogs with a humorous comment about the dog. As of February 2021, the twitter account has 8.9 million followers.

A tweet archive of the WeRateDogs twitter account was obtained and the dataset wrangled. The following sections describe the steps taken in the data wrangling process which was performed using a Jupyter notebook.

Gathering data

The project required three pieces of data to be gathered:

1. The WeRateDogs twitter archive. This was provided by Udacity as a csv file and manually downloaded.
2. The tweet image predictions ie - what breed of dog (or other object, animal) is present in each tweet according to a neural network. This file is hosted on the Udacity's servers and was downloaded programmatically.
3. Twitter API was queried for each tweet's JSON data. As I have not yet been granted access to a twitter developer account, I reviewed the twitter_api.py code provided by Udacity to understand how it works. Subsequent to this, I downloaded the tweet_json.txt file (also provided by Udacity).

Assessing data

After gathering each of the above pieces of data, they were assessed for:

- Quality eg: issues with content
- Tidiness eg: issues with structure that prevents easy analysis.

This was done using two types of assessment:

- Visual assessment - opening the csv files up in excel and identifying any quality or tidiness issues
- Programmatic assessment - using python to view various cuts and summaries of the data. Most commonly used functions included pandas.dataframe.info() to review for data types.

From my visual and programmatic assessment, i identified the following quality and tidiness issues:

- Quality:
    - Twitter_archive: 181 retweeted items
    - Twitter_archive: tweet_id is an integer but should be a string
    - Twitter_archive: timestamp is an object but should be in datetime format
    - Twitter_archive: rating_numerator varies a lot, some more than 10. Rating_numerator and rating_denominator are often within the text body so should be extracted and replaced.

    - Twitter_archive: rating_numerator and rating_denominator are of type int but should be float, to account for decimals.
    - Twitter_archive: some names of dogs are incomplete eg: missing names, or incorrect names (a, the)
    - Image_predictions: tweet_id is an integer but should be a string.
    - Image_predictions: remove 66 duplicated jpg_url rows.
    - Tweet_data: id column should be renamed to tweet_id
    - Tweet_data: tweet_id is an integer but should be a string.

- Tidiness:
    - Twitter_archive: Drop any columns that won't be used in the analysis
    - Twitter_archive: Split the timestamp column into two separate columns with the date and time
    - Twitter_archive: doggo, floofer, pupper and puppo should be values rather than column headings
    - Image_predictions: Drop any columns that won't be used in the analysis
    - Image_predictions: Inconsistent use of lower case and upper case for dog breed predictions
    - All tables should be merged into one.

Cleaning data

Before any cleaning was performed, copies of the original data were made and the cleaned data contained the suffix '_clean' at the end of each of the tables.

The cleaning process was split up into the following three steps - define, code and test.
I converted each of the quality and tidiness assessments into defined tasks, then converted those tasks to code and then tested the dataset to ensure that the cleaning operation worked.

One identified quality issue that was not cleaned was the dog names in the twitter_archive table. As noted above, there were a number of missing or incomplete names. On visually inspecting the data, I noted that the name was sometimes, but not always included in the text column. After consideration, I decided not to clean the 'name' column and to omit it from my analysis.