Wrangle & Analyze WeRateDogs Data
Wrangle Report

Introduction

The main goal of this project is to demonstrate the data wrangling abilities and skills those we learnt as part of the Udacity Data Analysis Nanodegree program.

In this data wrangling project, I will gather, assess and clean twitter posts from Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. This project will download 5000+ of their tweets as they stood on August 1, 2017.

Project details

The tasks of this project are as follows:

- **Gathering data**
- **Assessing data**
- **Cleaning data**

Data for the project was gathered from various sources:

- ### Twitter archive file

This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Udacity provided the twitter-archive-enhanced.csv file, which was manually downloaded and imported into a pandas data frame.

- ### The tweet image predictions

The highest ranked dog breed predictions for each dog image from the WeRateDogs Enhanced Twitter Archive are included in this file.
The Requests library is used to download data from a URL address and save it to a .tsv file. After that, the contents of the image-predictions .tsv file are loaded into a pandas data frame containing 2075 rows and 11 columns.
The highest ranked three forecasts, as well as the tweet ID, image URL, and image number for the most confident forecast, are all listed in the table.

- ### Twitter API File

The tweet id, favourite count, and retweet count are all contained in the Twitter API file. Udacity provided the data, which was manually downloaded and imported into a pandas data frame from the tweet-json.txt file. The dataframe has 2354 rows and two columns. The index is the tweeter ID column.

Assessing data

Following data collection, the data is evaluated for tidiness and quality as follows:

- ### Enhanced Twitter Archive
  - A sample of data is visually inspected first, with a summary of data types and non-null values shown. This makes it possible to spot columns with the wrong data type and/or null values. After that, duplicate IDs are checked.
  - Following that, the number of replies and retweets is determined.

  - The number of values in the name of dog column is validated programmatically.
  - Also, all tweets were examined for dogs who were assigned to more than one canine group (stage).
  - The denominator of the rating is visually examined by presenting a sample of data, and then ratings with a denominator greater than 10 are printed for further inquiry.
  - The numerator of the rating is also evaluated visually.
  - We examine the text column programmatically for any float ratings based on the visual assessment of rating columns.
  - Expanded URLs are visually evaluated first, and then programmatically checked for the presence of two or more URLs in a single column.

- ### Image Predictions
  - A set of data is visually inspected first, with a summary of data types and non-null values shown. This makes it possible to spot columns with the wrong data type and/or null values. The jpg url field was then examined for duplicates and to ensure that it only included jpg and png pictures. The 1st prediction is verified as the final stage to see how many images have been identified as dog images.

- ### Twitter API Data
  - A set of data is visually inspected first, with a summary of data types and non-null values shown. This makes it possible to spot columns with the wrong data type and/or null values. After that, duplicate IDs are checked.

Cleaning data

Pandas is used to tidy up the quality and tidiness issues noted in the Assessing Data section:

Enhanced Twitter Archive

- A copy of the dataset is made as a first step, and it will be used throughout the cleaning procedure. We eliminate these along with other unneeded columns because some of the retrieved tweets are replies and retweets.

- The four-column 'stage' classification for dogs (doggo, floofer, pupper, or puppo) has been consolidated into one.

- After that, we convert the timestamp to DateTime because it has the wrong data type (it's an object).

- Float ratings that were mistakenly read from the text of a tweet are collected once more, this time accurately. The fact that the rating numerators are bigger than the denominators does not require cleaning, but we introduce a normalised rating that will be utilised for graphs.

- We have 639 expanded URLs with several URL addresses; thus, we have used the tweet id field to create accurate links.

Image Predictions

- A copy of the dataset is made as a first step, and it will be used all throughout cleaning procedure. We renamed columns since some of the names were unclear and provided little information about the content.

- Now we tidy up the dog breeds by replacing underscores with whitespace and capitalising the initial letter to ensure that the formatting is uniform and clean.

- There were 66 image url duplicates deleted.

- We utilise the dog breed predicted in the 2nd or 3rd predictions for the remaining rows because only 2075 images have been categorised as dog images for the top prediction (1st prediction).

Twitter API Data

- In this data set, there is no need to conduct any cleaning tasks.
- As a final step in the cleaning process, we combine all datasets into one and export the results to the twitter archive master.csv file.