Wrangle Report

Udacity Data Analyst Nanodegree

Bilal Karim

Introduction

This project involves gathering data in various forms for the WeRateDogs Twitter handle. This page posts pictures of dogs that are sent to it, adding entertaining comments and ratings out of 10. The dogs are always given ratings above 10 (most commonly between 11 and 14).

Directions were provided to obtain the various data files that would be used in the project. First, the data was gathered into a Jupyter Notebook environment, then assessed both visually and programmatically to later be cleaned and analyzed. Details are provided below on each of the steps below, in the order the steps were taken.

Gathering Data

Information about the 3 files collected is provided below:

1) Twitter archive: This csv file was provided as "file on hand". It was downloaded and placed in the project directory.
2) Tweet image predictions: This .tsv file was downloaded programmatically from Udacity's servers into the Jupyter notebook.
3) Twitter retweet count and favorite count: This json data was acquired from the WeRateDogs Twitter handle using Twitter's API.

The files were placed in separate Pandas DataFrames for the next steps.

Assessing Data

The datasets were assessed for both quality (content) and tidiness (structure) using a mix of visual and programmatic assessment.

The methods I used to explore the data were: .sample(), .info().describe(), .duplicated(), and .sort_values(). I also narrowed down on specific records using .query and .loc.

I was required to identify and fix 8 issues with the data. The following issues were identified and fixed:

Data quality issues:
WeRateDogs Twitter archive

1) Columns have different value-counts. For example, expanded_urls column has 2297 records while others have 2356 records.

2) Since we are looking for original tweets only, the 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'in_reply_to_status_id' and 'in_reply_to_user_id' columns should be null.
3) There is no way to ensure that each tweet has an image attached. The image_pred file contains this information, and may be useful.
4) 'timestamp' is in string (object) format.
5) Since the ratings are all out 10 points, the 'rating_denominator' column is not required. Furthermore, some items have been read into ratings which are not meant to be so (for example, "4/20", "9/11", and "7/11").
6) Ratings numerator has values below 10 (which is incorrect based on the page's unique rating mechanism). Furthermore, ratings above 14 appear to be uncommon and possibly mistakes due to typing errors.
7) Name column has incorrect values. For instance, examples of some 'Name' column values with 'None': `tweet ID 667509364010450000` is Tickles, `tweet ID 668142349051129000` is Oliver, `tweet ID 677918531514703000` is Reese, but has an apostrophe.

Image Prediction file

8) There are duplicates in the jpg_url column.

Tidiness issues

1) WeRateDogs Twitter archive: Dog stages are classified in separate columns, breaking the 'each variable forms a column' rule of tidy data.
2) Twitter API data: This data is meaningless by itself. It can be combined with one of the other DataFrames to be add value to the dataset as a whole and for it to be considered tidy.

Cleaning Data

I created copies of the original DataFrames by appending '_clean' at the end of each of their names in order to keep the original DataFrames intact. I created a 'Data cleaning plan' that contained tasks to correct each of the issues I had identified in the previous step. Finally, I performed programmatic data cleaning for each of the issues identified in 3 stages:

1) Define: A summary of the cleaning activity to be done.
2) Code: Fix the quality or tidiness issue with the data, and then save it to the correct DataFrame.
3) Test: Ensuring that the code had worked as intended, and that it had not caused anything else to break.

One of the main lessons at this stage was the iterative nature of the data wrangling process. I had to revise some of the earlier steps in the process and identify and prioritize issues that could be fixed to satisfy the project motivations better.

The cleaned data was saved in a csv file named 'twitter_archive_master'.

Analyzing Data

Once the data was cleaned, I extracted 4 insights from the data:

1) Percentage of each rating_numerator in the data set
2) Times of day (hour:minute) for the top 10 posts by favorite_count
1) Relationship between retweet_count and favorite_count
2) Average rating numerator by dog stage