Wrangle Report

Wrangle & Analyze Data Project

Data Analyst for Shell Nanodegree

By

Jason Kanhai

September 7th 2021.

Introduction

In this brief report, I will outline the steps of the Data Wrangling processes employed in creating a dataset that is fit for analysis. The aim of this project was to wrangle data from the WeRateDogs Twitter account, gather supplemental data from additional sources, assess the quality and tidiness of the data and clean where necessary.

Gathering Data

For the purpose of analyzing data related to the WeRateDogs twitter account, data was acquired from three main sources:

- WeRateDogs Twitter Archive

This Twitter archive was already provided as part of the Udacity wrangling assignment in the form of a .csv file (twitter_archive.csv) which contained basic data on 2356 tweets sourced from the @weratedogs twitter account. This data was brought into the notebook using the pd.read_csv function as a 'twitter_archive' dataframe.

- Tweet Image Predictions

This dataset included image predictions of dog breeds for the dogs included in WeRateDogs tweet images. The data was acquired from Udacity's server using the Requests library and loaded to the Jupyter notebook.

- Tweet API & JSON

Each tweet's favourite and retweet count were acquired by querying the Twitter API for JSON data based on their individual tweet ID, using the Tweepy library. The JSON data was stored in a text file (tweet_json.txt_ that was then read into the notebook.

Assessing Data

The next step in the data wrangling process was done visually (with the aid of Microsoft Excel and Jupyter Notebooks) and programmatically using python code.

Issues identified were classified by Quality or Tidiness accordingly

Quality

Twitter Archive

1. Dataset contains retweeted ratings.
2. Lots of columns that do not add value to intended analysis including 'source', in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls'.
3. Some instances of inaccurate records of numerator and denominator.
4. 'rating_numerator' and rating_denominator' should be float to accommodate for decimal values.
5. Some records have large ratings due to rating of multiple dogs in an image.

6. 'timestamp should have date-time datatype.
7. Inaccurate 'name' recorded including 'a', 'an', 'such', 'the'.

Image Prediction

1. Duplicate URLs exist

Tweet_json

1. Retweets exist in dataframe

Tidiness

1. 'doggo', 'floofer', 'pupper' and 'puppo' columns in the tweet_archive dataframe are cumbersome and should technically be condensed into a single column describe dog type.
2. Too many columns on image prediction and confidence levels in image_prediction dataframe, should be diluted to one
3. 'tweet_id' datatype not consistent across 3 datasets, must be converted to int in tweet_json.
4. Datasets should be merged into a single dataframe for analysis

Cleaning Data

The final stage of the data wrangling process involved defining the actions needed to rectifiy the identified issues, write, and execute the code that would fix them and test to ensure its success. Cleaning steps taken in this project include

1. Removing retweets from twitter_archive
2. Removing columns that were not necessary for intended analysis from twitter_archive
3. Melting dog classification columns into a single dog_type column
4. Converting rating_numerator and rating_denominator to float data type to accommodate decimal values.
5. Replacing numerators and denominators with accurate values where errors occurred.
6. Account for large numerators and denominators by creating a new measure of rating (numerator/denominator)
7. Convert timestamp from object to date-time
8. Replace inaccurate names with 'none'; too many names to replace.
9. Remove duplicate 'jpg_url' from image_prediction.
10. Condense image predictions into single dog_breed column
11. Remove retweets from tweet_json
12. Convert tweet_id in tweet_json to integer to facilitate joining
13. Join all datasets into a single dataframe for analysis.

Conclusion

By following these three key steps of the Data Wrangling process, data is now fit for use in conducting further analyses that would provide more accurate insights than if used raw.