We Rate Dogs Wrangle Report

AUTHOR: JAJA TAMUNO-OMI

Introduction:

As part of the data analyst Nanodegree program and in completion of the data wrangling lesson, I was tasked with wrangling, analysing, and finally visualizing data 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. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage. The steps taken are written in detail below:

Gathering Data:

Here we gather the three key pieces of data for the project namely

- **WeRateDogs twitter archive**: Provided by Udacity in a flat file containing tweets and other information like, text, source and individual dog ratings, there were 5000+ tweets originally which were filtered to contain only original tweets and no retweets
- **Image predictions files:** This file contains breed predictions for the pictures found in **WeRateDogs** dog rating tweets, it also includes the prediction confidence of the model for each image, the data was gather using the request library to get its content via the provided url, which was then written into an image-predictions.tsv file.
- **Additional data using Tweepy API:** Using the keys provided by my twitter developer account, I gathered additional data (Retweets and favourite_count) for tweets in the original archive file via tweepy, although some tweets in the original archive were no longer available

Assessing Data:

In this step I assessed the three datasets for quality and tidiness issues both programmatically and visually , the issues are listed as follows:

Quality issues:

1. Some tweets have repeated links in the expanded_urls field
2. Tweet_id, image number is an integer, Retweeted_status_id, retweeted_status_user_id, in_reply_to_status_id, in_reply_to_user_id, ratings numerator & denominator are float, timestamp is an object
3. Source column in archive_df contains html anchor tag.
4. Seperate timestamp column into various date-based columns i.e year, month, day.
5. The column names in images predictions aren't descriptive i.e p1, p1conf, p1_dog.
6. Some tweets are retweets and not original ratings.
7. Some Predicted dog breed names in predictions contains hyphens.
8. Drop unused columns.

Tidiness issues:

1. doggo, floofer, pupper, puppo should be in a single column dog stage.
2. The archive tweets, image predictions and additional tweet data tables should be combined into one for this project

Cleaning Data

In this section I remedy the issues identified in the earlier stage, I start by creating copies of the datasets to work with so that I can always have the original to work with late, I then follow the steps of stating the issue, defining how I will clean the issue, converting my definitions into executable code and finally testing the data to ensure the code was implemented correctly, I performed this process for each of the 10 issues identified in the assessing stage

Storing analysing and visualizing data

At this point I stored the now cleaned and combined dataset, for analysis and visualization, by identifying trends and presenting them with visualizations like charts.

Conclusion

During the Data Wrangling and Analysis project, I had the opportunity to use many libraries like Pandas, Numpy, Tweepy, request etc, this enabled me to gather, access and clean WeRateDogs twitter archive and finally provide insights on the data.