DATA ANALYTICS NANODEGREE PROGRAM (UDACITY)

WeRateDogs Twitter Archive Wrangling Report

Osigah Ogedegbe | September 2022

INTRODUCTION

In this report I outline the efforts to gather, clean and analyze over 5000 tweets from a twitter account @dogrates – WeRateDogs- and draw some insights from them. Weratedogs is a twitter account that rates people's dogs with a funny comment about the dog. The account is quite popular among users.

The data wrangling process is in 3 phases:

1. Data Gathering
2. Data assessment
3. Data cleaning

Data Gathering

I gathered data from 3 sources, stored in separate files:

- WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
- The image predictions file, programmatically downloaded from the Udacity servers
- The Json text in the lesson because I was not granted elevated access to the twitter developer account so I could not query the tweepy api. The **tweet_id**, **favourite_count** and **retweet_count** were extracted programmatically from this file as per requirements.

I loaded the 3 raw data files into separate tables: **archive_df**, **df_imgpred** and **df_tweets** collected for further assessment and cleaning.

Assessment and Cleaning

I started the assessment with the **archive_df** table and visually assessed the data which I realized contained some data quality and assessment issues such as inconsistency of name starting with both lowercase and uppercase letters, some dog names starting with “a” and some missing values.

After which I did a programmatic assessment using standard python methods such as info(), describe(), value_counts(), etc which revealed a lot more than the visual assessment.

The following problems were discovered and cleaned

- Numerator and Denominator values are not standardized and was cleaned by dropping a denominator values less than 10 and then dropping it entirely, retaining only numerators less than 15 and then making it one column called rating.
- Source in the archive dataset was too long and was split to only show the devices
- Retweet columns: retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and reply columns: in_reply_to_status_id, in_reply_to_user_id. (181 and 78 respectively) were not needed in the analysis and so the rows were first dropped then followed by the columns
- Timestamp Column is in object/ string format and was converted to datetime
- The expanded_urls column contains some missing data and was dropped
- The Name Column contains some invalid names that begin with lowercase letters (109 in total). They were dropped
- The Name column also contains 745 dog names with "None". There were changed to NaN values.
- The dog names and predictions are not consistent as not all began with capital letters which was eventually cleaned
- There are duplicated image urls in the image prediction dataset which was eventually dropped
- The archive table dog classifications, puppo, floofer, doggo and pupper were concatenated and the ‘None’ values were changed to NaN.
- The predictions (img_pred) table itself was not cleaned. There were many tweets with no dog breed predicted, these were left as is. The best prediction for breed and associated confidence level were extracted and merged into the archive table.
- The json_data (tweet_df) table itself was not cleaned. The retweet_count and favorite_count columns were merged into the archive table.
- The remaining cleaned columns numbering 1854 entries in the archive table were then saved to the new “twitter_archive_master.csv” file.