Data Wrangling Report (WeRateDogs)

By Elliot Manuel Sithole

06/09/2022

Context

The project involves the wrangling of data from various sources associated with tweets from Twitter; WeRateDogs. This rates pictures of dogs in a humorous manner, most often giving ratings higher than 10/10.

The Objectives

The main objective of this project is to perform data wrangling that is gathering, assessing and cleaning from three different sources. The focus is on improving and fixing quality and tidiness issues and then storing, visualising and analysing the wrangled data to review vital insights.

Step 1:

In this project, all three pieces of data were sourced from three different means and uploaded into the pandas data frame:

1. Twitter archive file: manual direct download of the file (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv),

2. The tweet image predictions: the dog breed prediction is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and downloaded programmatically using the Requests library and URL link (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv).

3. Twitter API & JSON: Data gathered for each tweet's retweet count and favourite count at a minimum. Using the tweet IDs in the WeRateDogs Twitter archive, to query the Twitter API for each tweet's JSON data using Python's *Tweepy library* and to store each tweet's entire set of JSON data in a file (*tweet_json.txt file*). Each tweet's JSON data is written to its line and then read as .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favourite count.

Step 2:

Quality

After a thorough visual and programmatic assessment, several issues were documented, about quality and tidiness shortfalls.

Source File name	**Quality Issue**	**Action taken**
archive	Wrong data type in the *TimeStamp* column data.	Change column type to datetime.
archive, tweet, image prediction	Wrong data type in the *tweet_id* columns data from all data frames.	Change all tweet_id columns to strings using astype pandas function
archive	Unnecessary columns with very low imputed data; retweeted_status_id, in_reply_to_status_id,in_reply_to_user_id,retweeted_status_timestamp and retweeted_status_user_id.	Drop columns in the archive with unnecessary features: drop pandas function
archive	Values in the rating denominator are not equal to 10.	Check and assign so that all rating denominators should be equal to 10: use query and assign functions
archive	Values in the rating numerator column with errors and are unrealistic	Visually inspect extraction from texts errors: use excel after using the query function to filter data
archive	Some dog names are missing, unidentified or weird.	*Replace missing, unidentified or weird names with* none: use replace pandas function
images	Image predictions that are not dogs.	To drop, use the query function that meets; the condition that accepts any dog

		prediction that is only dogs(True)
archive	unrated dogs in the data set.	Drop rows with unrated dogs. Use the drop function

Tidiness

Source File name	Quality Issue	Action taken
images	Inconsistent uppercase and lower case letters on dog types in columns p1, p2 and p3.	Make dog type names p1,p2 and p3 consistent using the *str.title()* method.
tweet	Inconsistent naming of id column from both tweet and archive data frames.	Use *rename* function to change *id* column name in tweet to *tweet_id*
archive	The dog stage (doggo, floofer, pupper, puppo) columns in archive data frame can be condensed into a single column.¶	Create a single column dog_stage from doggo, floofer, pupper, and puppo columns and drop these columns
All	Combine all 3 dataframes into one.	Merge all 3 dataframes for the dog data into a single dataframe on *tweet_id*: use pd.merge, how Left function

Results

The final file; save gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv" ready for further analysis and visualizations.